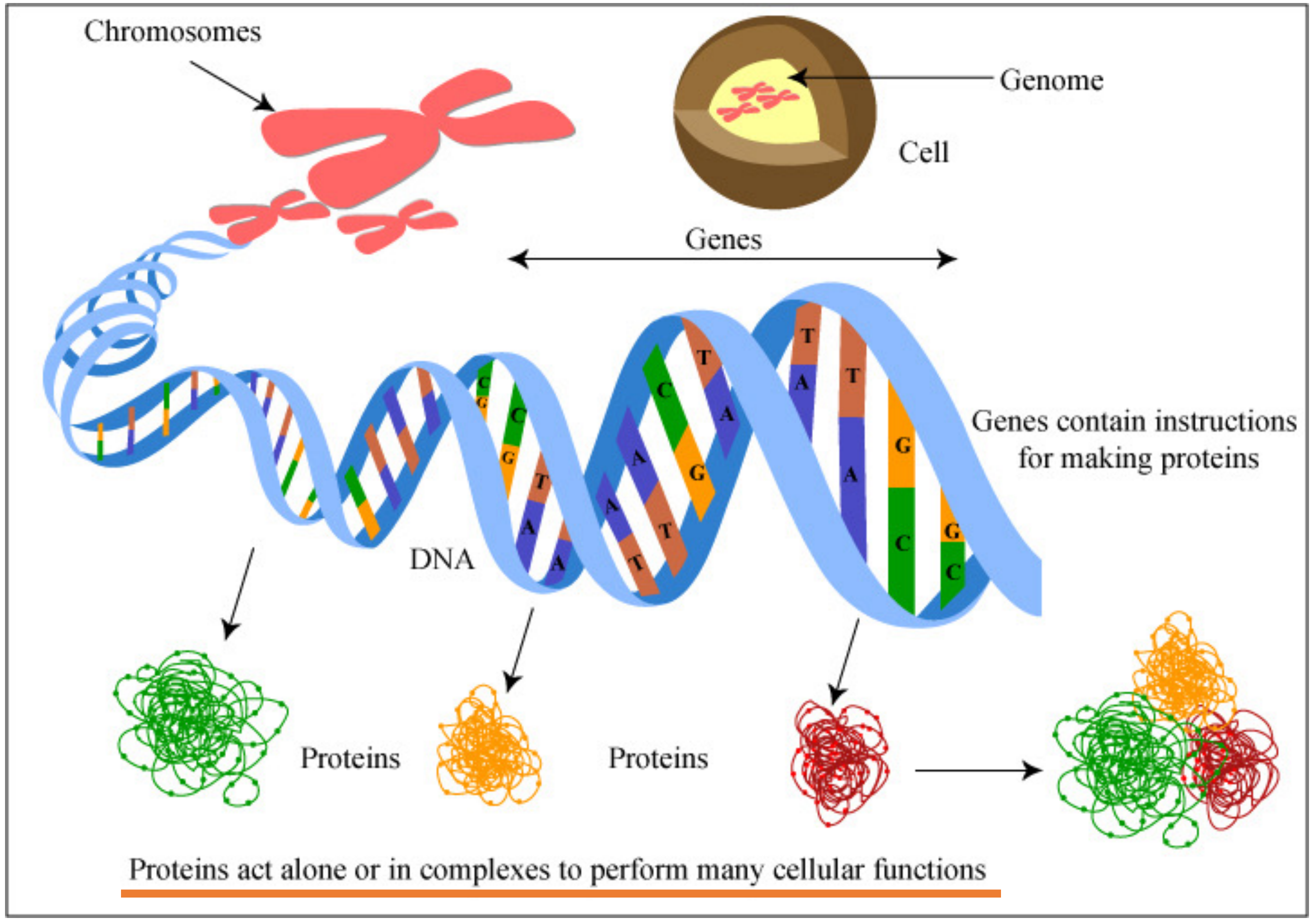
Algorithms for understanding the spatial and network organization of biological systems

Uthsav Chitra

March 1, 2024

PRINCETON
UNIVERSITY

Biological systems are organized across a hierarchy of scales: from genes and proteins...



Chromosomes
Genome
Cell
Genes
Genes contain instructions
for making proteins
DNA
Proteins
Proteins
T
A
T
C
A
G
A
G
C
A
T
G
A
G
C
T
A
A
T
G
C
C
G
G
Proteins act alone or in complexes to perform many cellular functions

Image from MIT OpenCourseWare

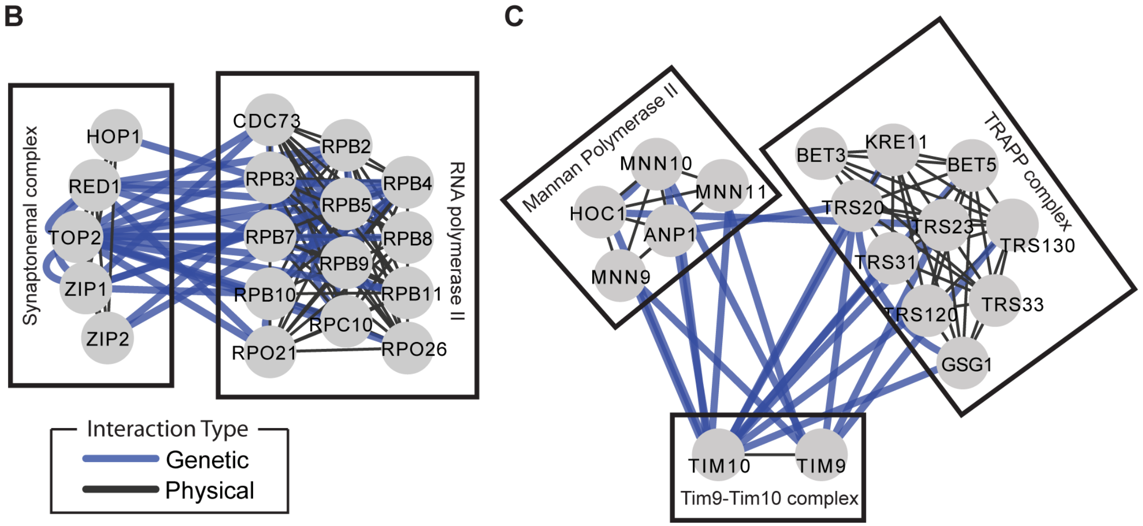


Image from Hannum et al., PLOS Genetics 2009

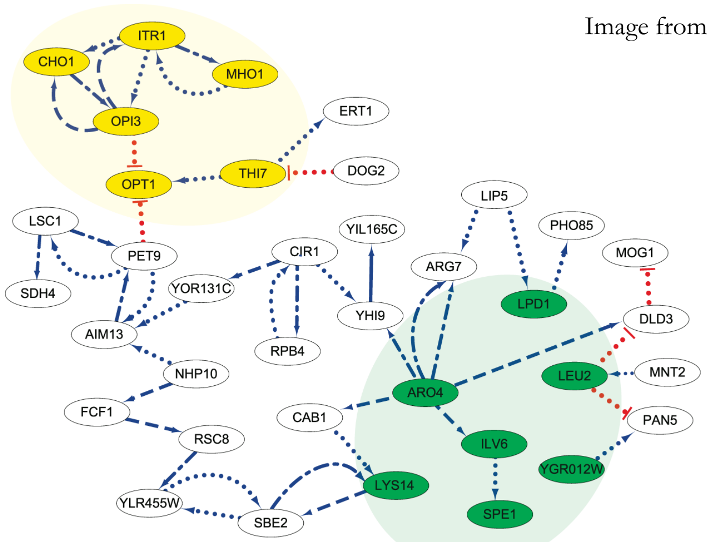
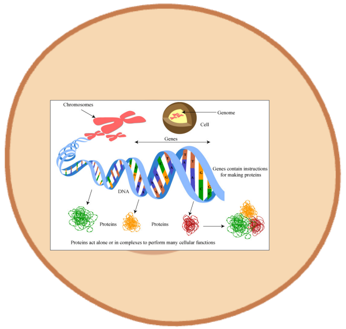


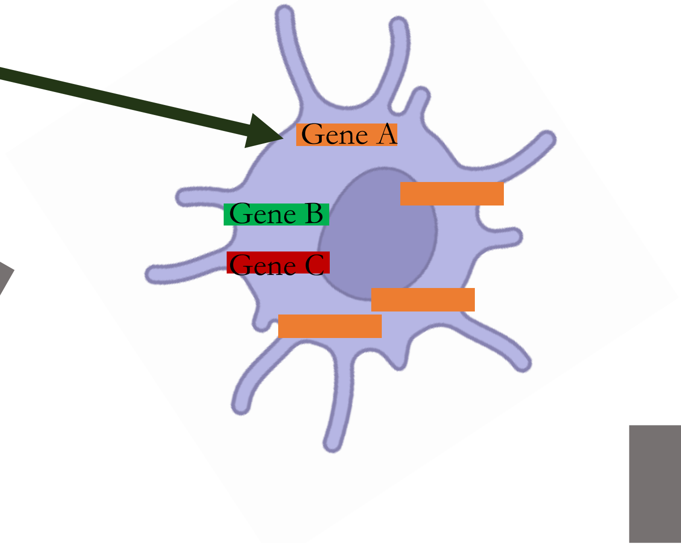
Image from Chen et al, Scientific Reports 2019

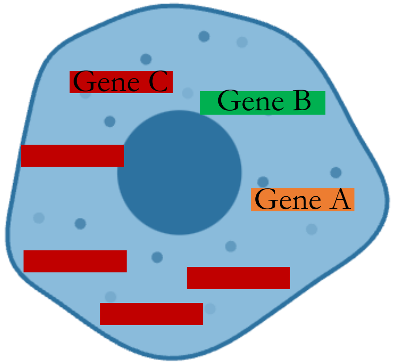
…to cells and tissues

Gene expression
(e.g. mRNA transcript)

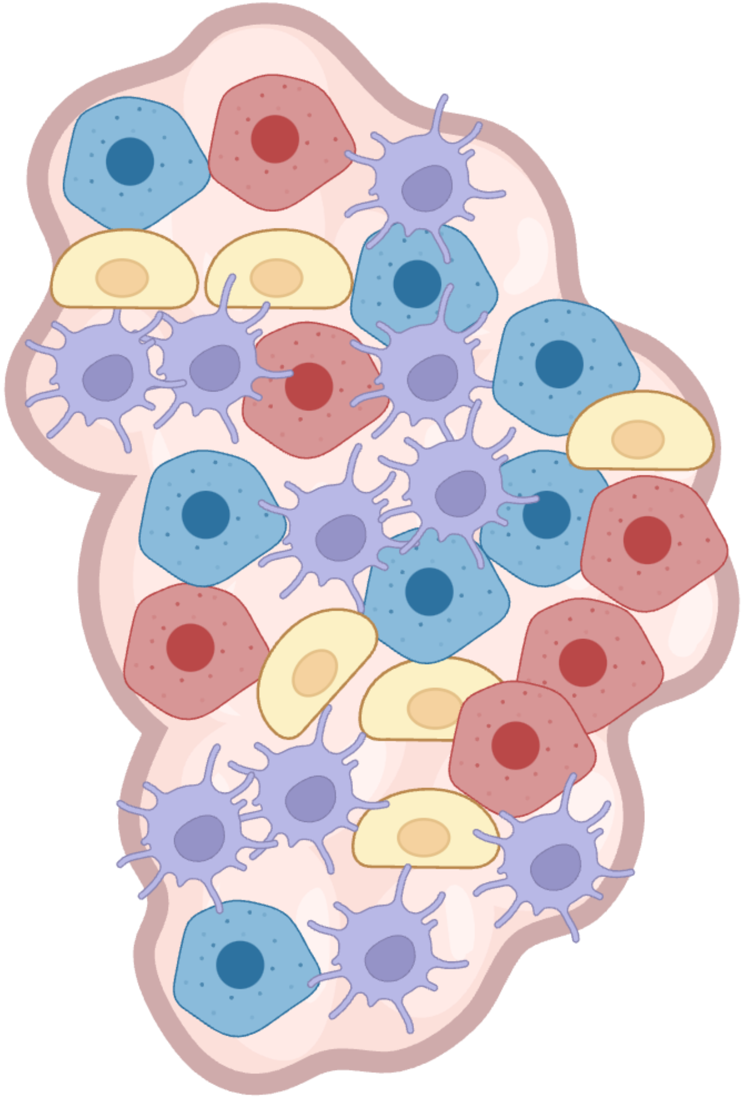


Fibroblast
(High expression of Gene A)



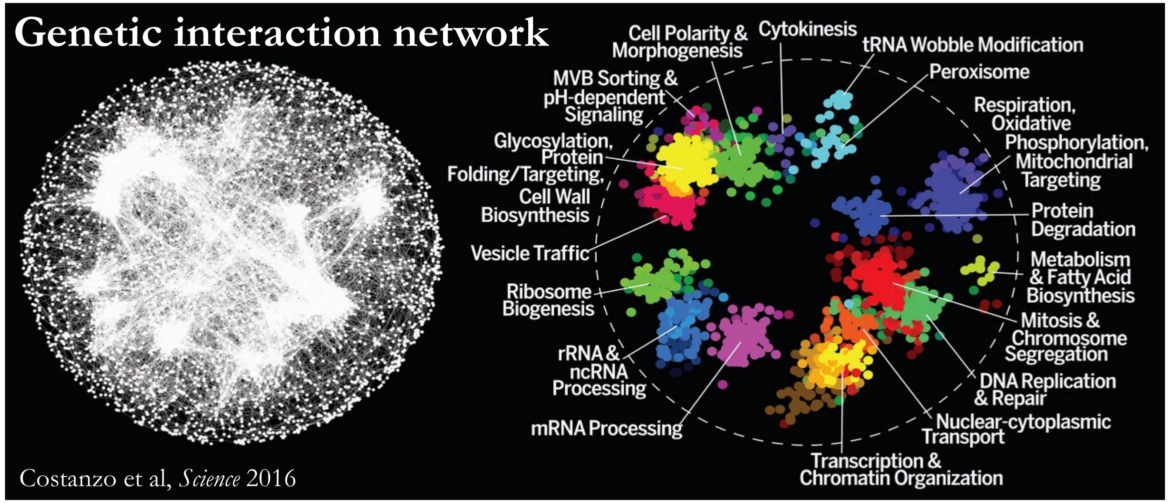


T-cell
(High expression of Gene C)

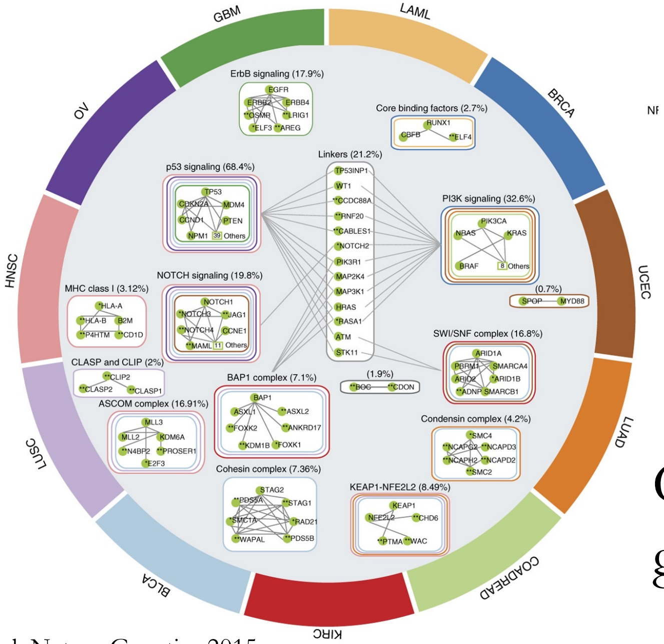


Spatial arrangement of cells in tissue

Genetic interactions and cellular organization impact human health and disease

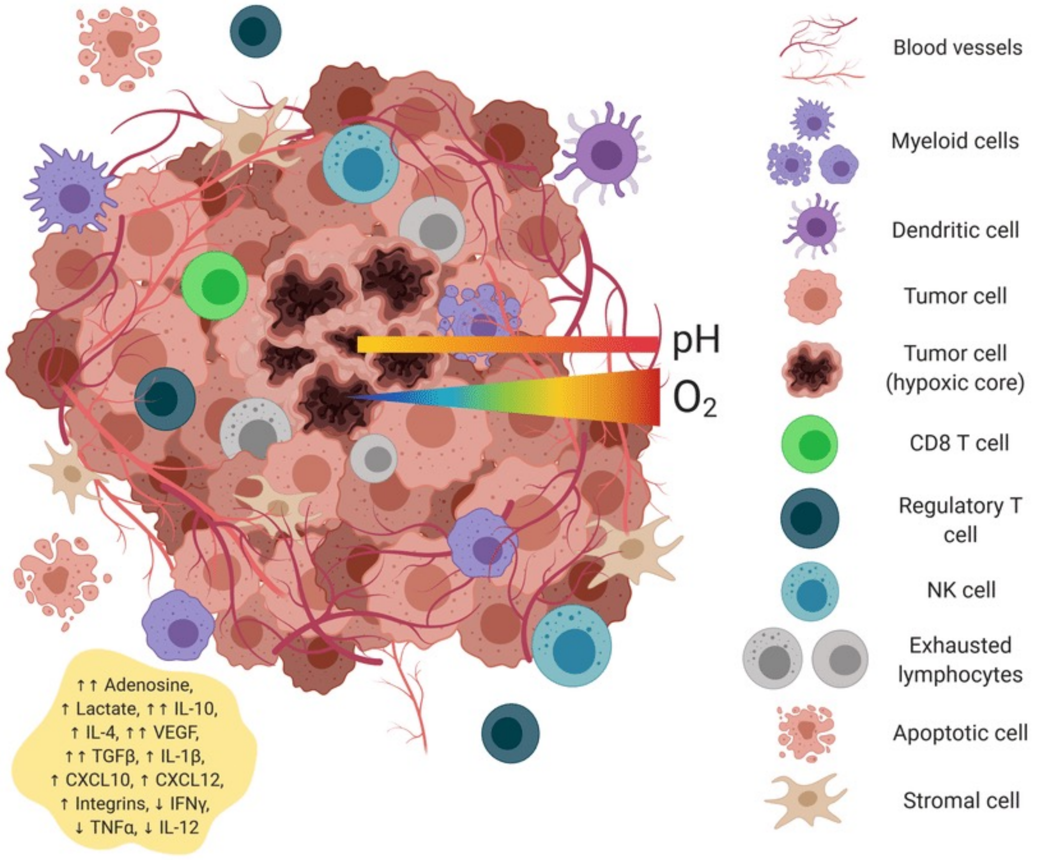

Genetic interaction network
Cell Polarity & Morphogenesis
Cytokinesis
tRNA Wobble Modification
Peroxisome
MVB Sorting & pH-dependent Signaling
Respiration, Oxidative Phosphorylation, Mitochondrial Targeting
Glycosylation, Protein Folding/Targeting, Cell Wall Biosynthesis
Protein Degradation
Vesicle Traffic
Metabolism & Fatty Acid Biosynthesis
Ribosome Biogenesis
Mitosis & Chromosome Segregation
rRNA & ncRNA Processing
DNA Replication & Repair
mRNA Processing
Nuclear-cytoplasmic Transport
Transcription & Chromatin Organization

Costanzo et al, *Science* 2016



Leiserson et al, Nature Genetics 2015

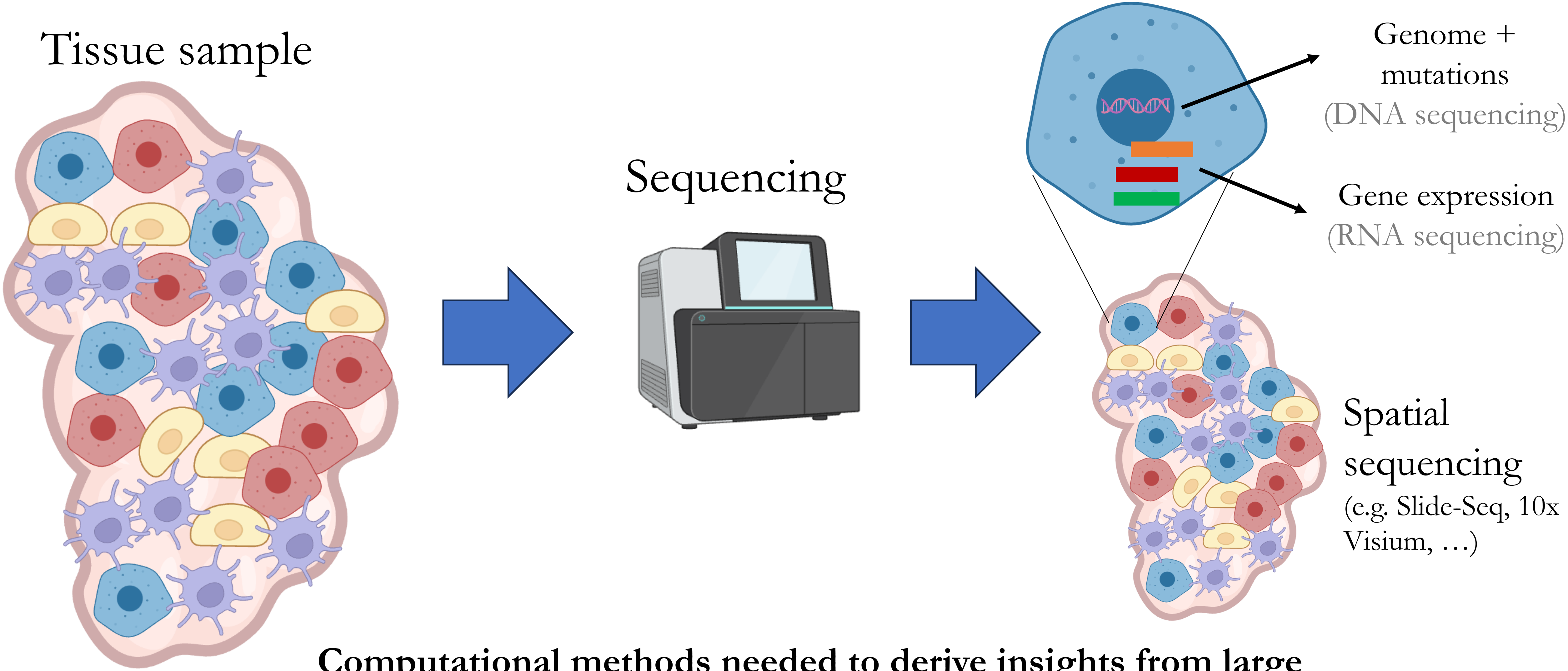
Cancer mutations alter gene networks



Fernandez et al, Int J Mol Sci 2019

Spatial heterogeneity in the tumor microenvironment

High-throughput sequencing data enables study of biological systems



Tissue sample

Sequencing

Genome + mutations

(DNA sequencing)

Gene expression

(RNA sequencing)

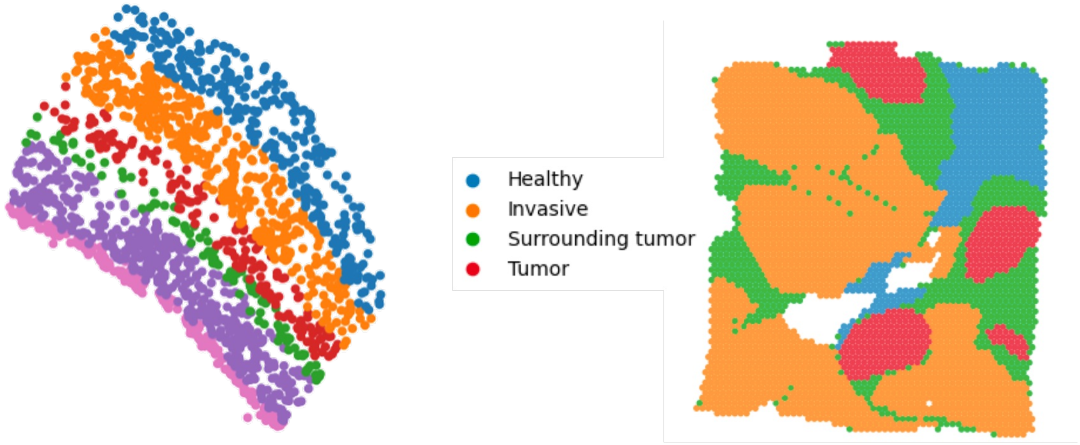
Spatial sequencing

(e.g. Slide-Seq, 10x Visium, …)

Computational methods needed to derive insights from large volume of sequencing data

My thesis: computational methods for understanding complex biological systems

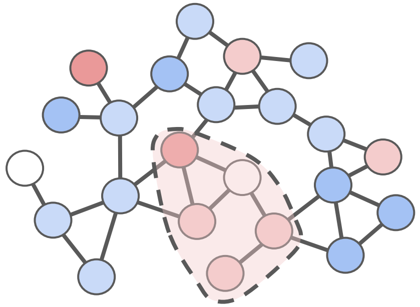
Spatial biology

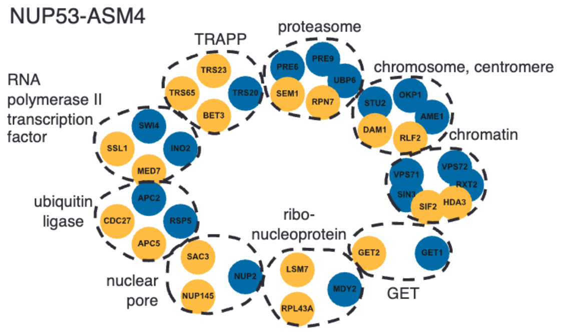


Spatial variation in gene expression

- Ma*, **Chitra***, et al. *RECOMB 2022 + Cell Systems.*
- **Chitra** et al. *RECOMB 2024 + in review at Nature Methods.*

Network interactions and anomalies





Altered subnetwork identification

- Reyna*, **Chitra***, et al. *RECOMB 2020 + JCB.*
- **Chitra** et al. *ICML 2021.*
- **Chitra***, Park*, Raphael. *RECOMB 2022 + JCB.*

Cell-cell interactions

- Sarkar*, **Chitra***, et al. *In submission at ISMB 2024.*

Learning genetic interactions

- **Chitra***, Arnold*, Raphael. *In review at Nature Genetics.*
- Shuaibi*, **Chitra***, Raphael. *In submission at RECOMB-CCB.*

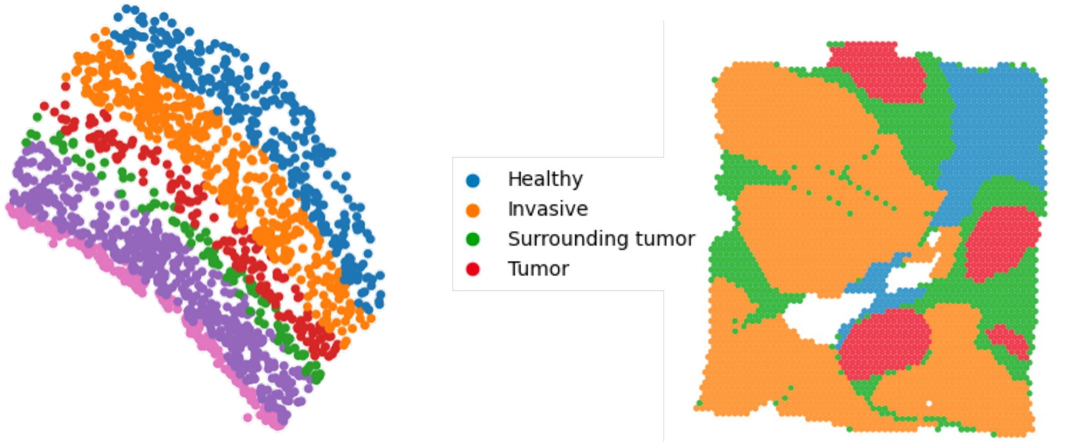
Machine learning + data mining

- **Chitra** and Raphael. *ICML 2019.*
- **Chitra** and Musco. *WSDM 2020.*

* indicates joint first authorship

My thesis: computational methods for understanding complex biological systems

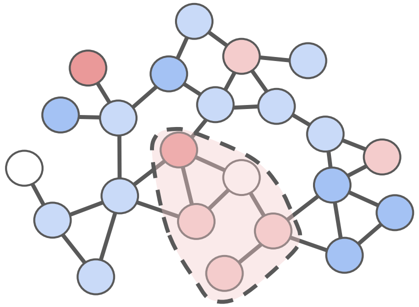
Spatial biology

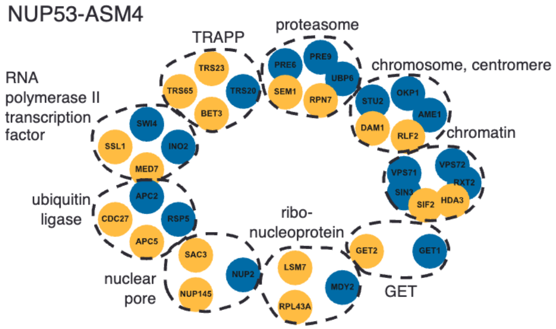


Spatial variation in gene expression

- Ma*, **Chitra***, et al. *RECOMB 2022 + Cell Systems.*
- **Chitra** et al. *RECOMB 2024 + in review at Nature Methods.*

Network interactions and anomalies





Altered subnetwork identification

- Reyna*, **Chitra***, et al. *RECOMB 2020 + JCB.*
- **Chitra** et al. *ICML 2021.*
- **Chitra***, Park*, Raphael. *RECOMB 2022 + JCB.*

Cell-cell interactions

- Sarkar*, **Chitra***, et al. *In submission at ISMB 2024.*

Learning genetic interactions

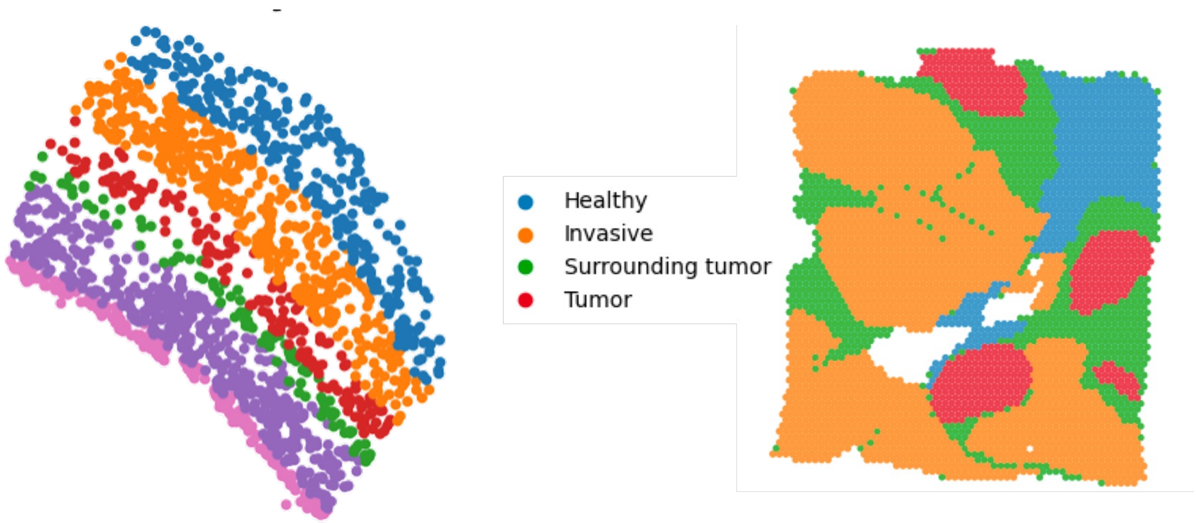
- **Chitra***, Arnold*, Raphael. *In review at Nature Genetics.*
- Shuaibi*, **Chitra***, Raphael. *In submission at RECOMB-CCB.*

Machine learning + data mining

- **Chitra** and Raphael. *ICML 2019.*
- **Chitra** and Musco. *WSDM 2020.*

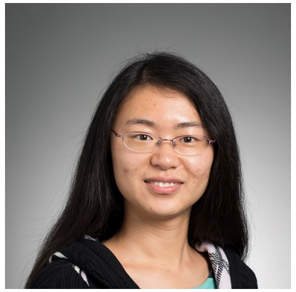
* indicates joint first authorship 🟨 : new work since pre-FPO

Modeling spatial variation in gene expression



Ma*, **Chitra*,** Zhang, Raphael. *RECOMB 2022 + Cell Systems.*

Chitra et al. *RECOMB 2024 + in review at Nature Methods.*



Cong Ma

Shirley Zhang

Brian Arnold

Hirak Sarkar

Sereno Lopez-Darwin

Kohei Sanno

Ben Raphael

* indicates joint first authorship

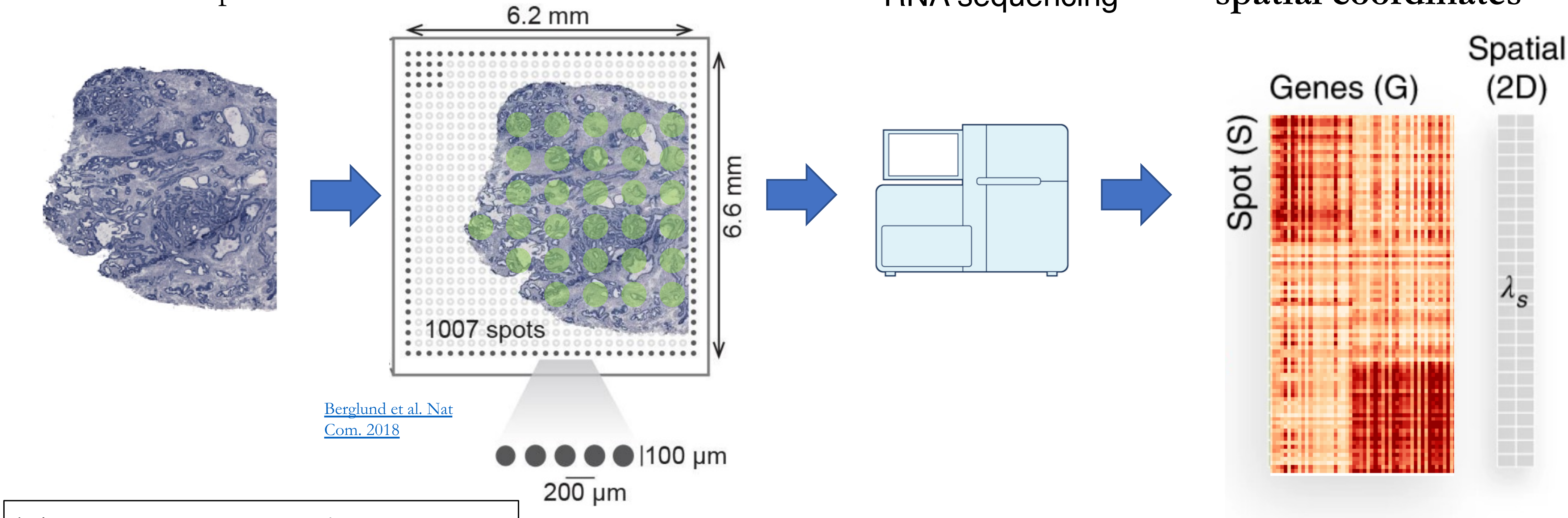
Spatially Resolved Transcriptomics (SRT)

Tissue sample

Barcoded Grid of Spots

RNA sequencing

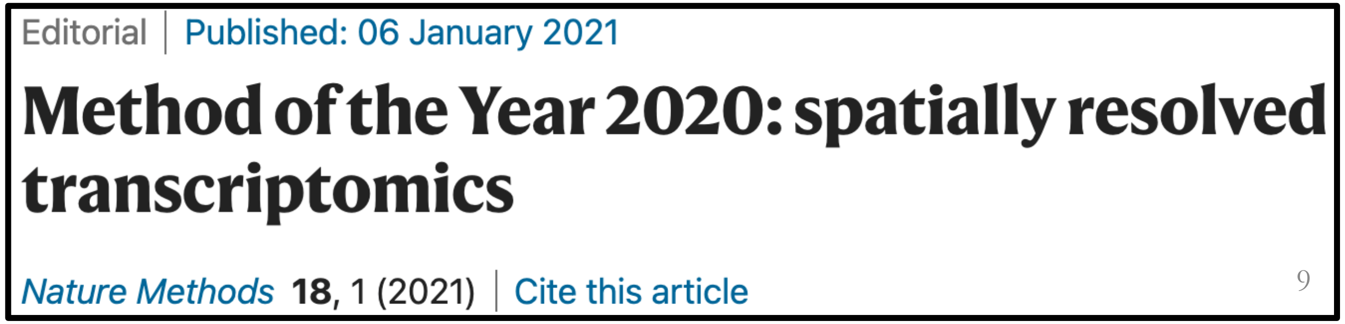
Gene expression matrix + spatial coordinates



Berglund et al. Nat Com. 2018

Technologies: Slide-Seq, 10x Visium, MERFISH, …

High-throughput: measure 1,000-20,000 genes at 1,000-10,000 spatial locations (each spot contains 1-20 cells)


Editorial | Published: 06 January 2021

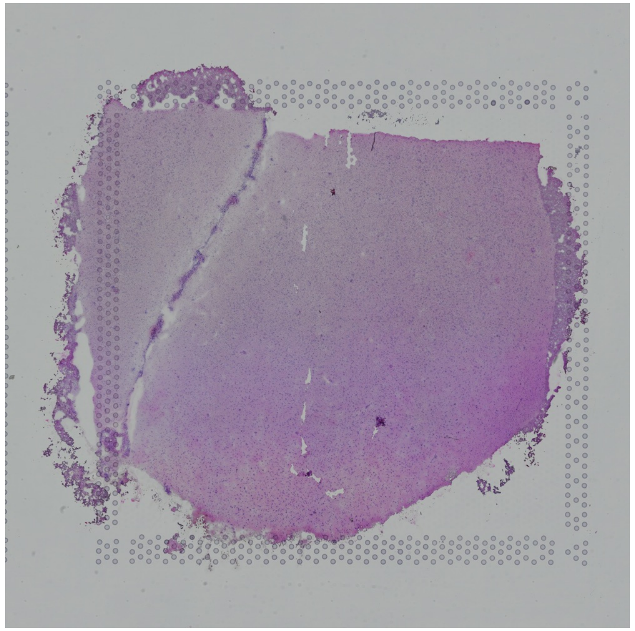
Method of the Year 2020: spatially resolved transcriptomics

Nature Methods **18**, 1 (2021) | Cite this article

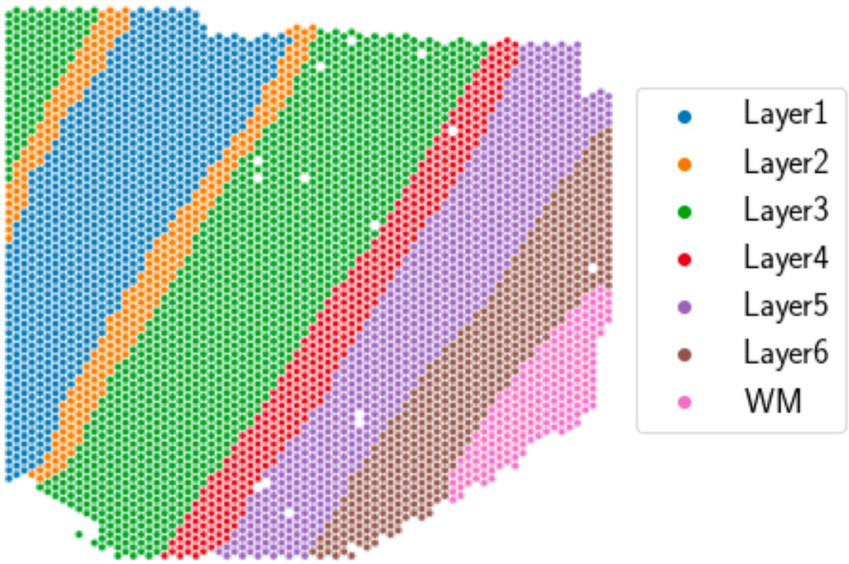
Spatially resolved transcriptomics (SRT) reveals new biology

Spatial domains/cell types

H&E stain



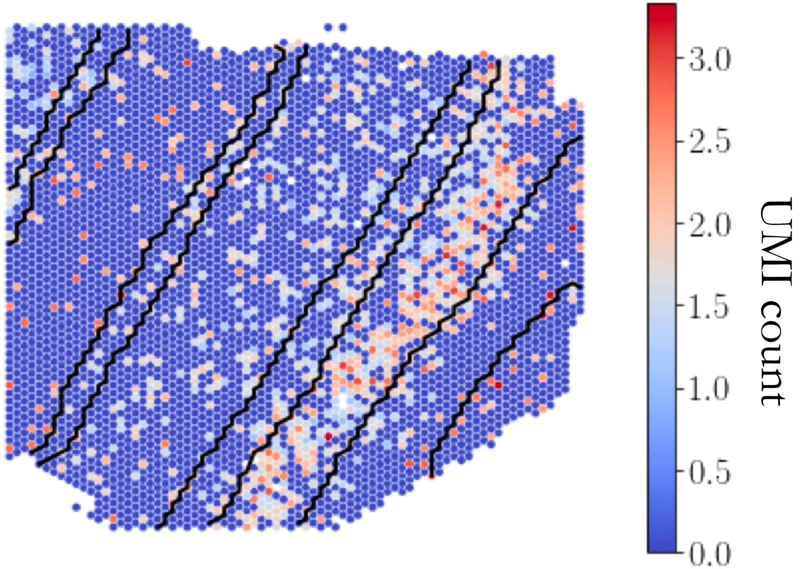
Neuronal tissue layers



Marker genes

(differentially expressed across domains)

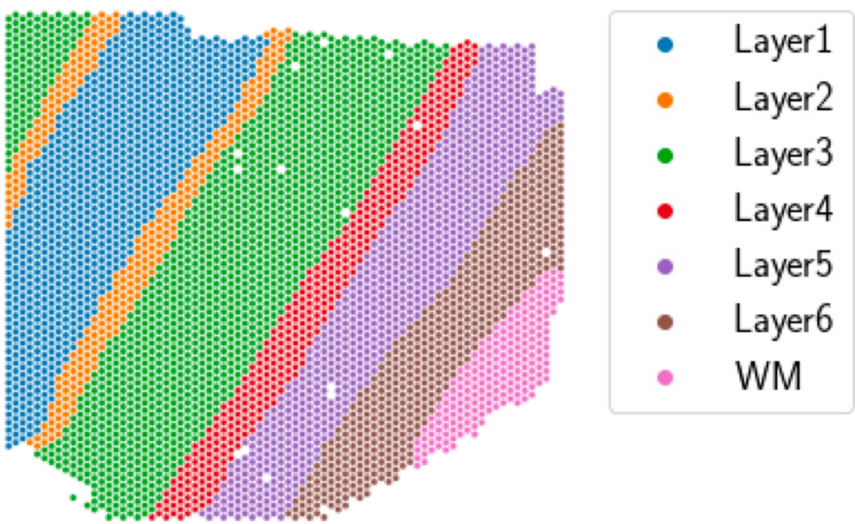
Layer5 marker gene *PCP4*



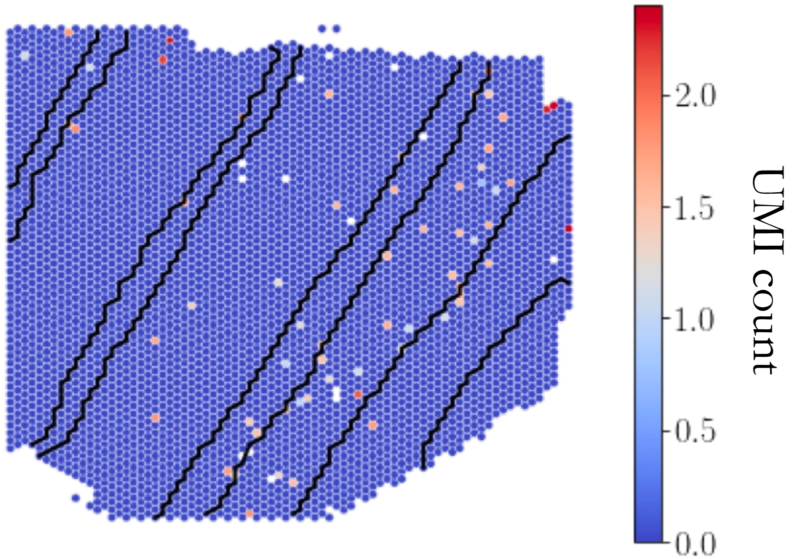
Human dorsolateral pre-frontal cortex (DLPFC) [Maynard et al., *Nat Neurosci* 2021]

Challenge: SRT data is **very sparse**!

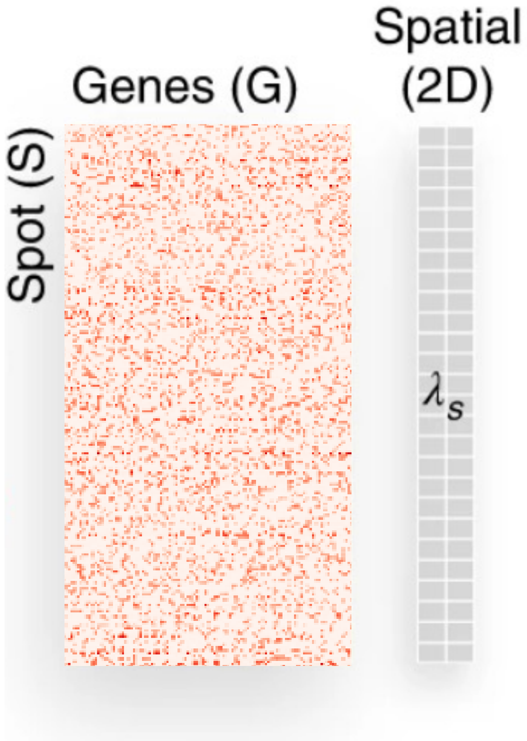
Sample 151508



Sparse expression of marker gene *TRABD2A*



Median gene has non-zero expression in <5% spots

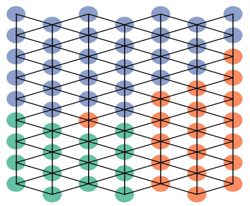


Sparse matrix: >90% zeros

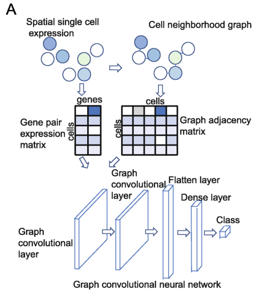
Overcoming sparsity by incorporating spatial information

Most algorithms use **local models**: nearby spots have similar cell type / expression

- **Hidden Markov Random Field (HMRF)**: BayesSpace [Nat. Biotech 2021], SPICEMIX [Nature Genetics 2022], Giotto [Genome Biology 2021] , scGCO [Nat Comm 2022] …
- **Graph neural networks (GNN)**: SpaGCN [Nature Genetics 2021], STAGATE [Nature Communications 2022] , SEDR [Genome Med 2024] ,…
- **Gaussian Processes**: SpatialDE [Nature Methods 2018], SPARK [Nature Methods 2020], SPARK-X Genome Biology 2021] , nnSVG [Nat Comm 2023] …

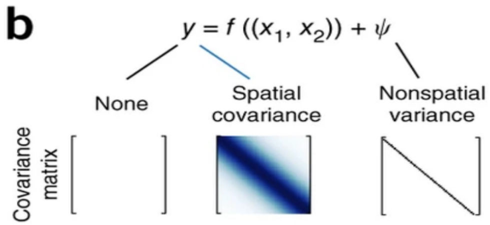


Hidden Markov Random Fields


A
Spatial single cell expression
Cell neighborhood graph
genes
cells
Gene pair expression matrix
cells
cells
Graph adjacency matrix
Graph convolutional layer
Flatten layer
Dense layer
Class
Graph convolutional layer
Graph convolutional neural network

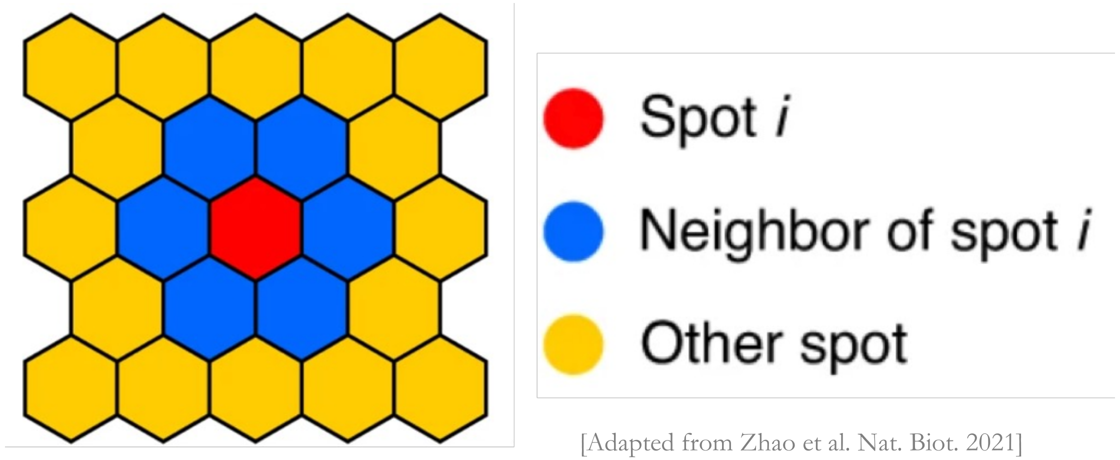
Yuan and Bar-Joseph, Genome Biol. 2020

Graph Neural Networks



Svensson *et al.*, Nat. Methods 2018

Gaussian Processes



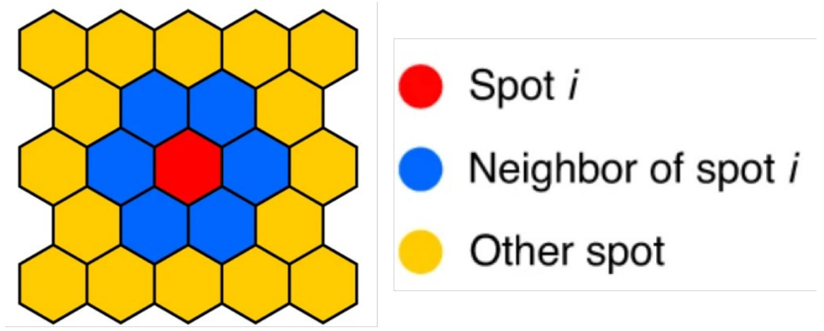
[Adapted from Zhao et al. Nat. Biot. 2021]

Overcoming sparsity by incorporating spatial information

Most algorithms use **local models**: nearby spots have similar cell type / expression

- **Hidden Markov Random Field (HMRF)**: BayesSpace [Nat. Biotech 2021], SPICEMIX [Nature Genetics 2022], Giotto [Genome Biology 2021] , scGCO [Nat Comm 2022] …
- **Graph neural networks (GNN)**: SpaGCN [Nature Genetics 2021], STAGATE [Nature Communications 2022] , SEDR [Genome Med 2024] ,…
- **Gaussian Processes**: SpatialDE [Nature Methods 2018], SPARK [Nature Methods 2020], SPARK-X Genome Biology 2021] , nnSVG [Nat Comm 2023] …

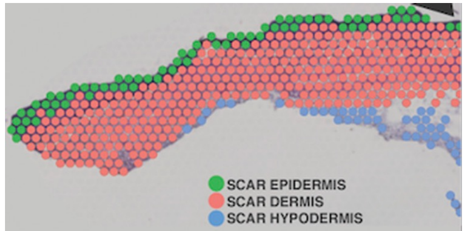
Local spatial information



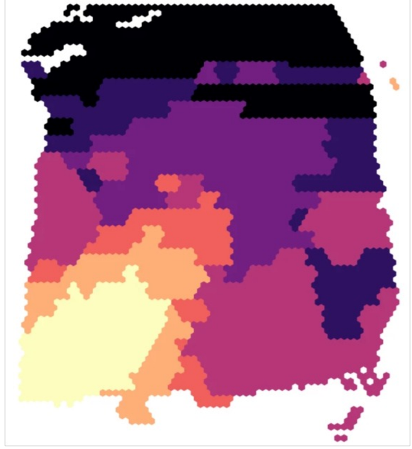
[Adapted from Zhao et al. Nat. Biot. 2021]

<u>Global model:</u> Cortex is made of layers!

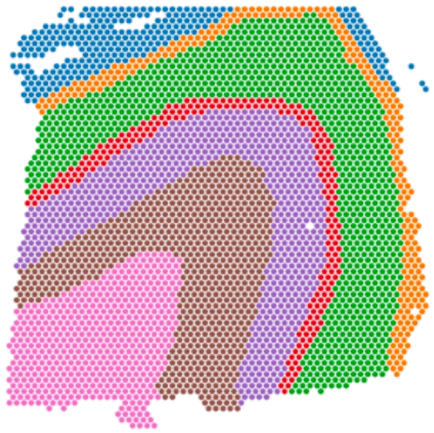
- Many layered tissues: skin, ureter, eye, …
- Can we incorporate the <u>layered geometry</u> in a gene expression model?



Skin
(Foster et al., 2021)



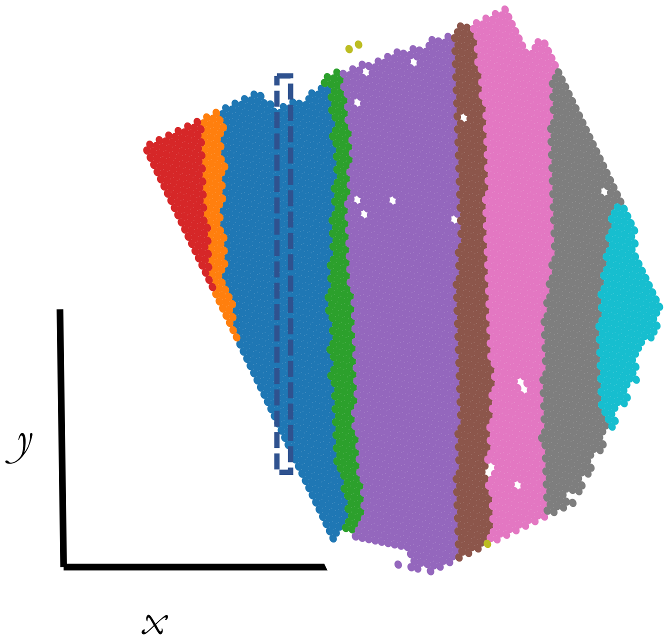
Spatial domains learned by Giotto (local model)



Annotated layers

A simple layered tissue

DLPFC sample 151508
(approximately axis aligned)

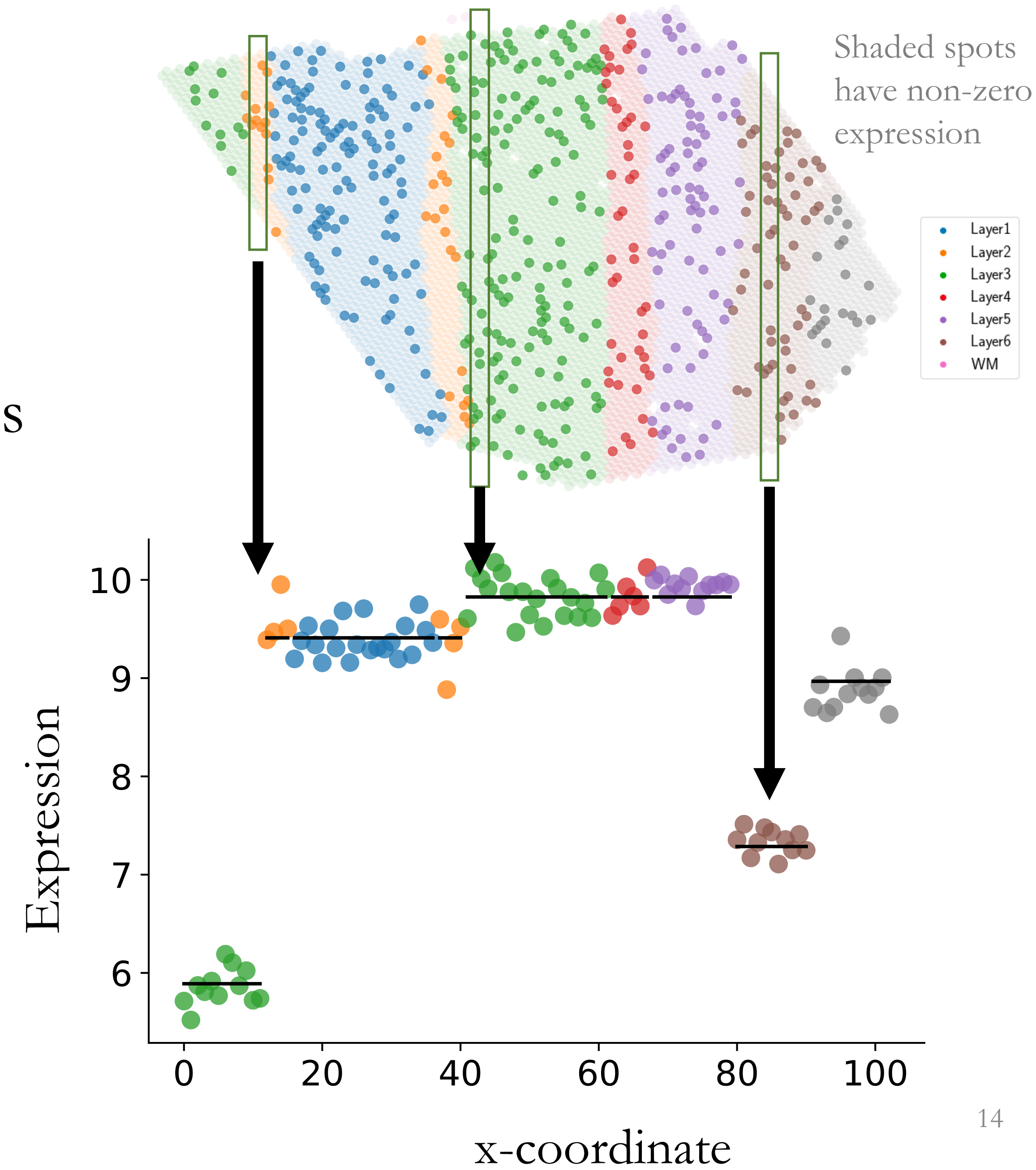


$f_g(x, y) = f_g(x)$ is piecewise constant

(Marker) gene expression is
≈***constant*** along *y*-axis

expression only depends on x-coord = distance to layer boundary (layer depth)

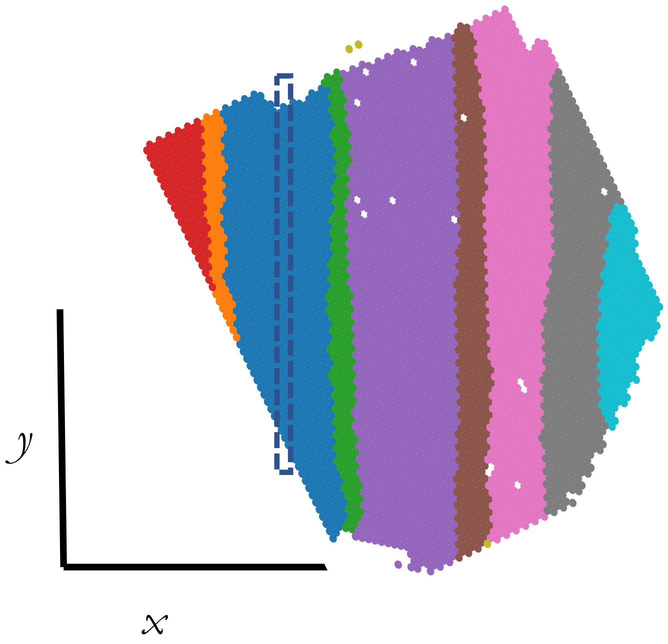
Pool sparse expression along y-axis



Shaded spots have non-zero expression

A simple layered tissue

DLPFC sample 151508
(approximately axis aligned)

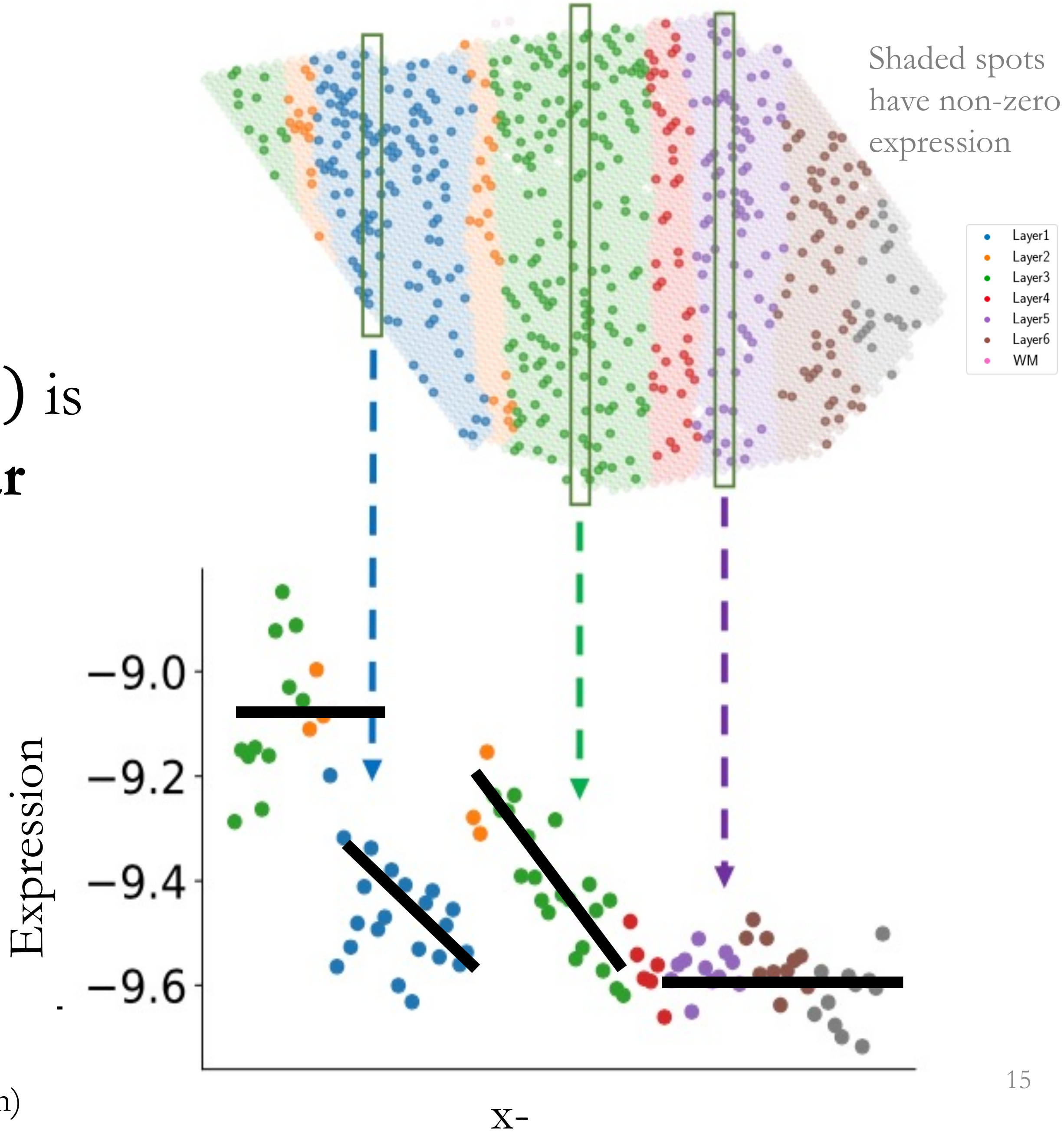


$f_g(x, y) = f_g(x)$ is **piecewise linear**

(Marker) gene expression is
≈***constant*** along y-axis

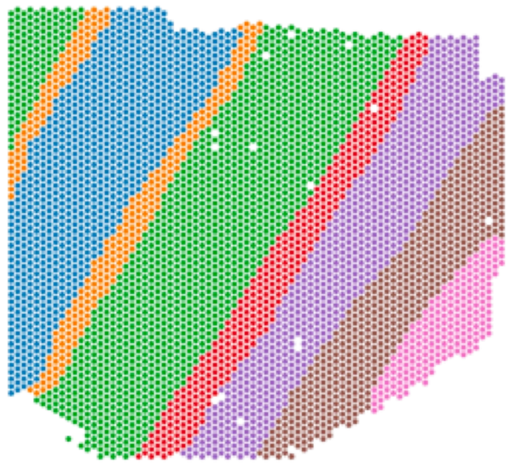
expression only depends on x-coord = distance to layer boundary (layer depth)

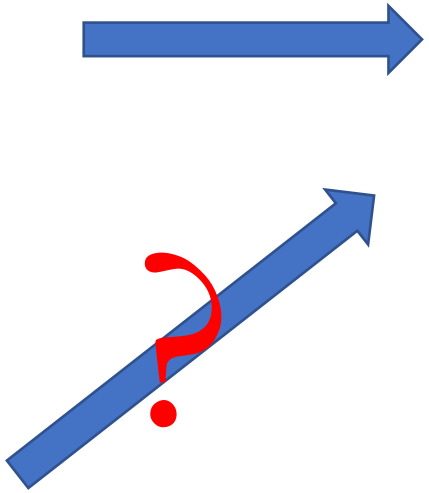
Pool sparse expression along y-axis

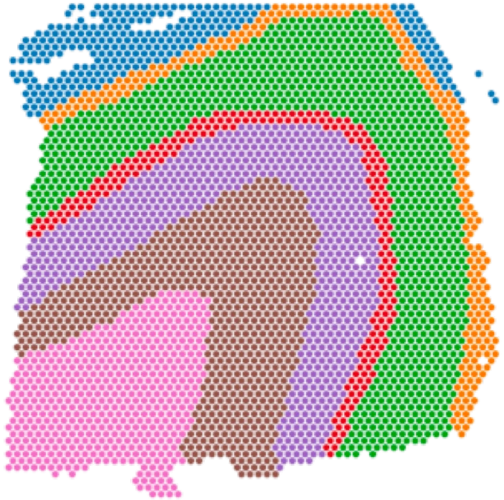


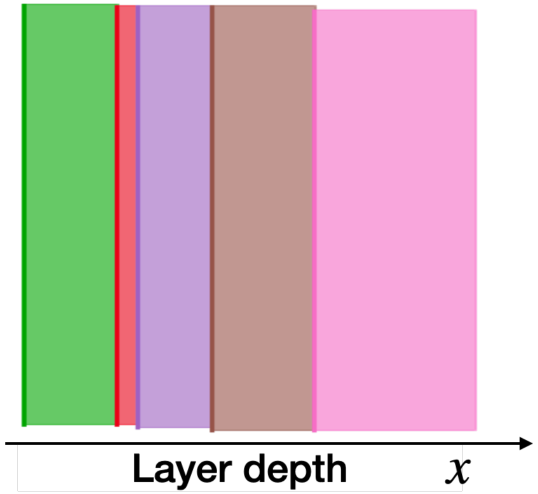
Shaded spots have non-zero expression

How to model layer depth in tissues with complex layered geometry?

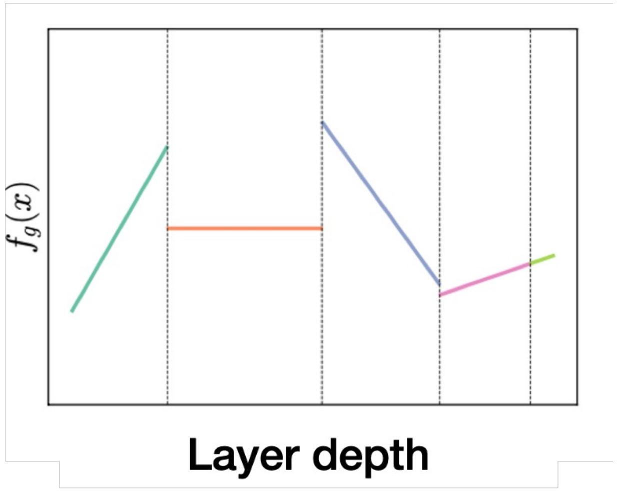






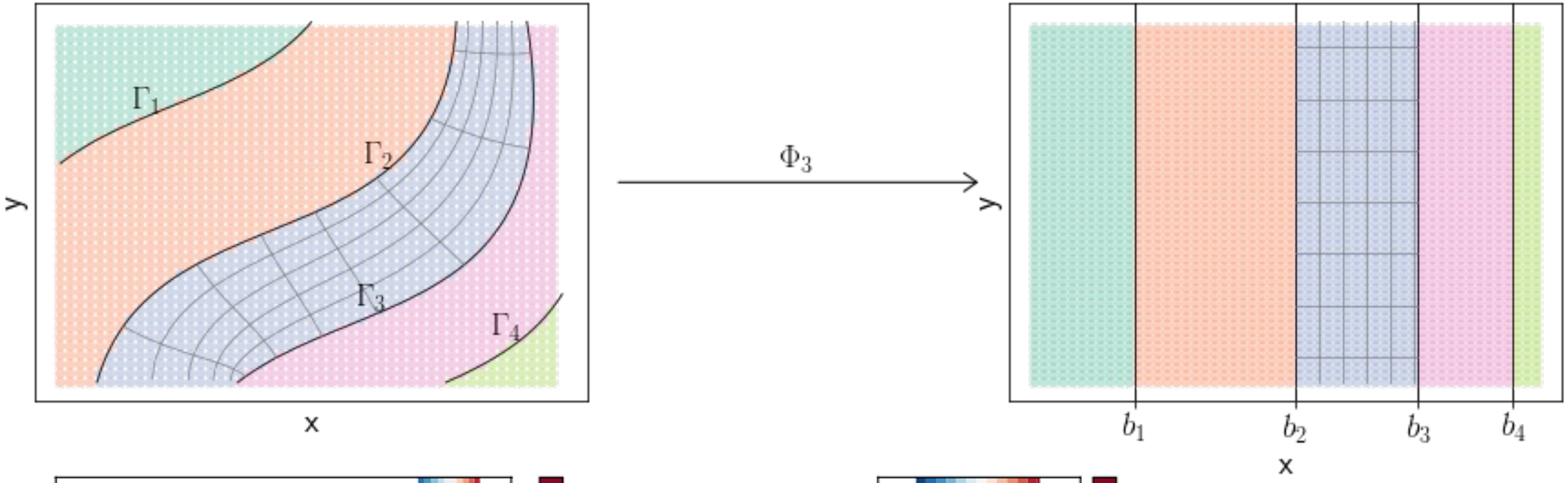


Global layered tissue geometry

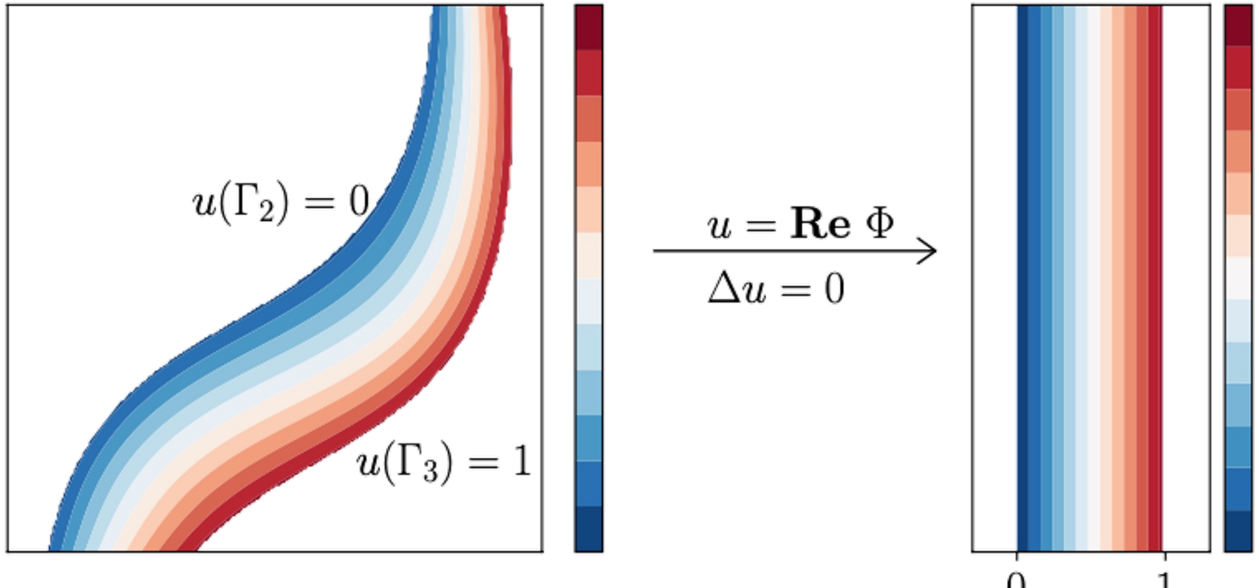


Expression function: Piecewise linear function of layer depth

Conformal maps and harmonic functions model layer depth



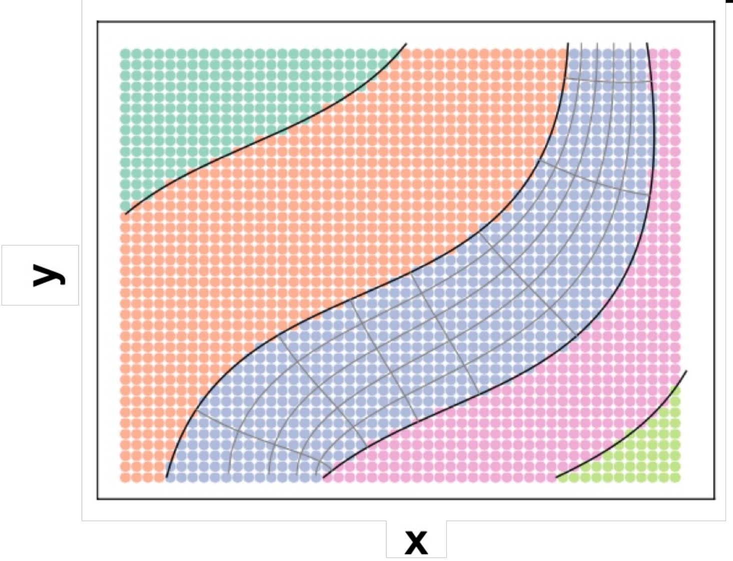
A **conformal map** $\Phi\colon D \subseteq \mathbb{C} \to \mathbb{C}$ **locally preserves angles** between curves

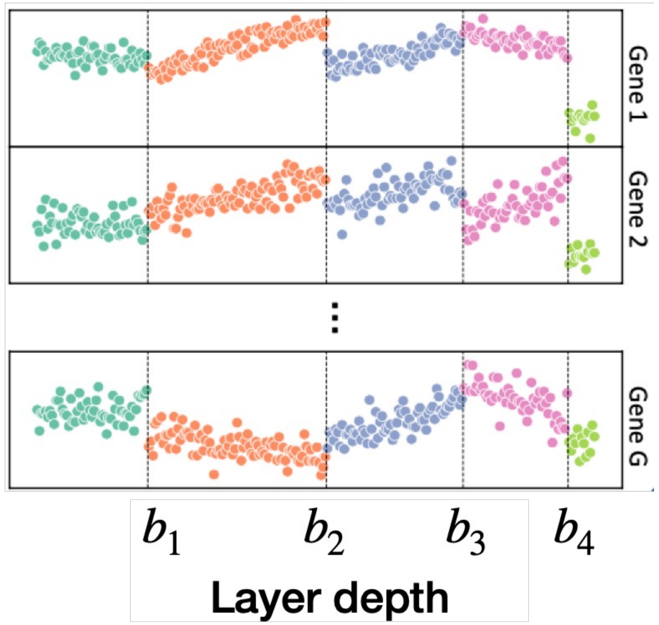


u = Re Φ is a **harmonic function** (satisfies heat eq)

Layer depth = **isotherms** (contours) of heat equation

Layered tissue problem formulation





Input

- Spot coordinate $\mathbf{s}_i = (x_i, y_i)$.
- Transcript count matrix $A = [a_{i,g}]$, $a_{i,g}$ for i^{th} spot and g^{th} gene.
- Number of layers L.

Probabilistic model

$$a_{i,g} \sim \text{Poisson}\Big(C_i \exp(f_g(\mathbf{\Phi}(x_i, y_i))\Big)$$

Piecewise linear expression: f_g

Global layered geometry: layer boundaries at $b_1, b_2, \ldots, b_{L-1}$

Maximum likelihood objective

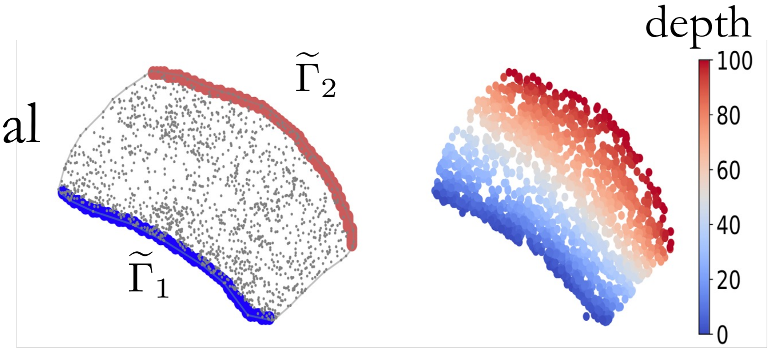
$$\underset{\substack{\text{breakpoints } b_1<b_2<\cdots<b_{L-1} \\ \text{piecewise linear } f_1,\ldots,f_G \\ \text{conformal maps } \mathbf{\Phi}=(\Phi_1,\ldots,\Phi_L)}}{\arg\max} \sum_{g=1}^{G} \left(\sum_{i=1}^{\text{N}} \log P\big(a_{i,g} \mid f_g(\mathbf{\Phi}(x_i, y_i))\big) \right)$$

Solutions to special cases

$$\underset{\substack{\text{breakpoints } b_1<b_2<\cdots<b_{L-1}\\ \text{piecewise linear } f_1,\dots,f_G\\ \text{conformal maps } \boldsymbol{\Phi}=(\Phi_1,\dots,\Phi_L)}}{\arg\max} \sum_{g=1}^{G}\left(\sum_{i=1}^{N}\log P\big(a_{i,g}\mid f_g(\boldsymbol{\Phi}(x_i,y_i))\big)\right)$$

Case 1: Approximate layer boundaries $\widetilde{\Gamma}_i$ are given

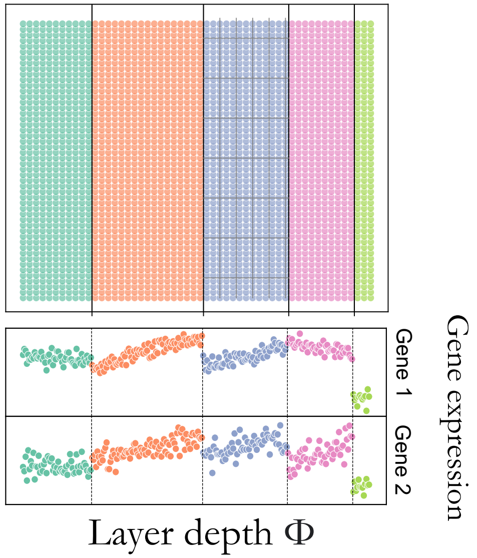
Step 1: Construct conformal map(s) Φ by solving heat equation



Step 2: Segmented regression

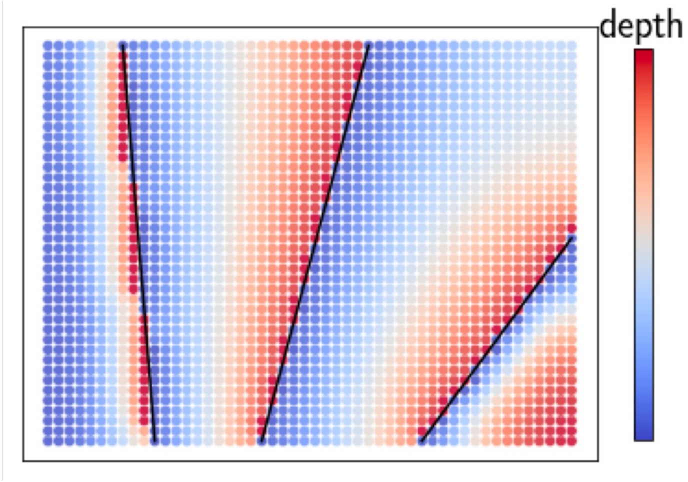
[Bai and Perron, *Econometrica* 1998]

- dynamic programming algorithm in O(LN^{2}G) time
- L=#layers, N=#spots, G=#genes

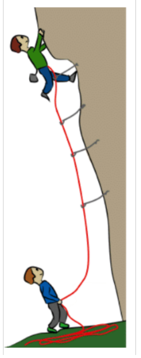


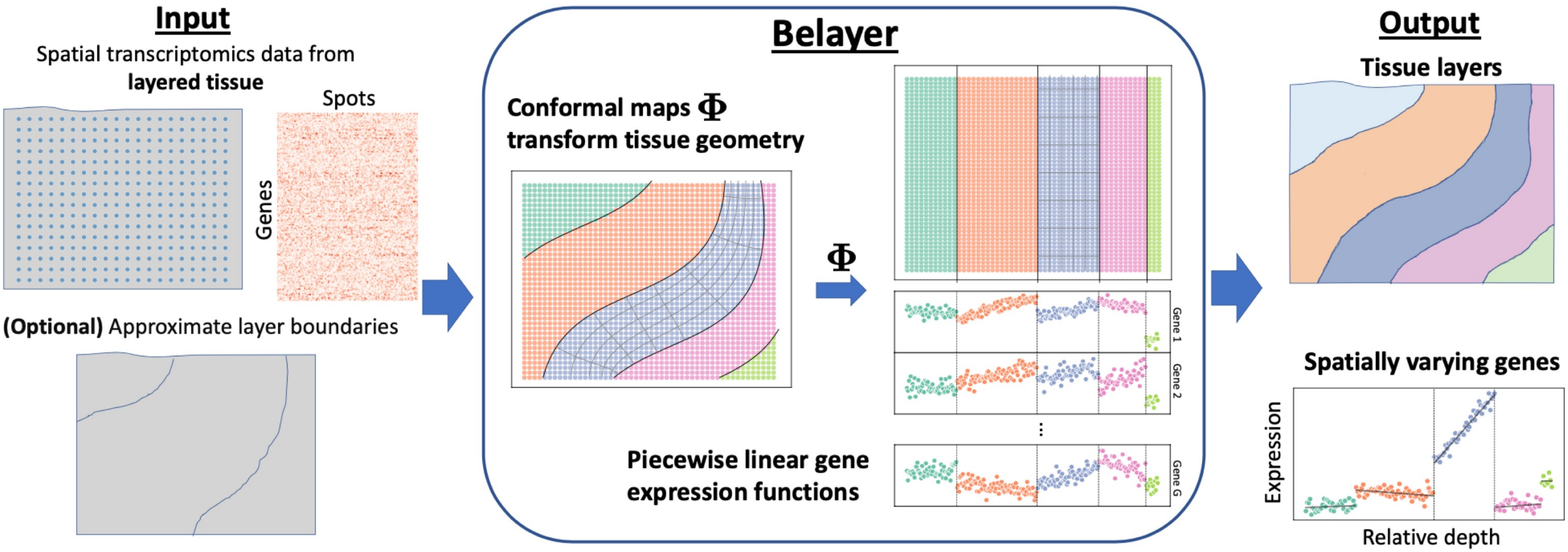
Case 2: Layer boundaries Γ_i are non-intersecting lines (not given)

DP algorithm (~Nussinov algorithm for RNA folding) to find best lines Γ_i, conformal maps $\boldsymbol{\Phi}$, piecewise functions f_g



Overview of Belayer





Input

Spatial transcriptomics data from **layered tissue**

(Optional) Approximate layer boundaries

Belayer

Conformal maps Φ transform tissue geometry

Piecewise linear gene expression functions

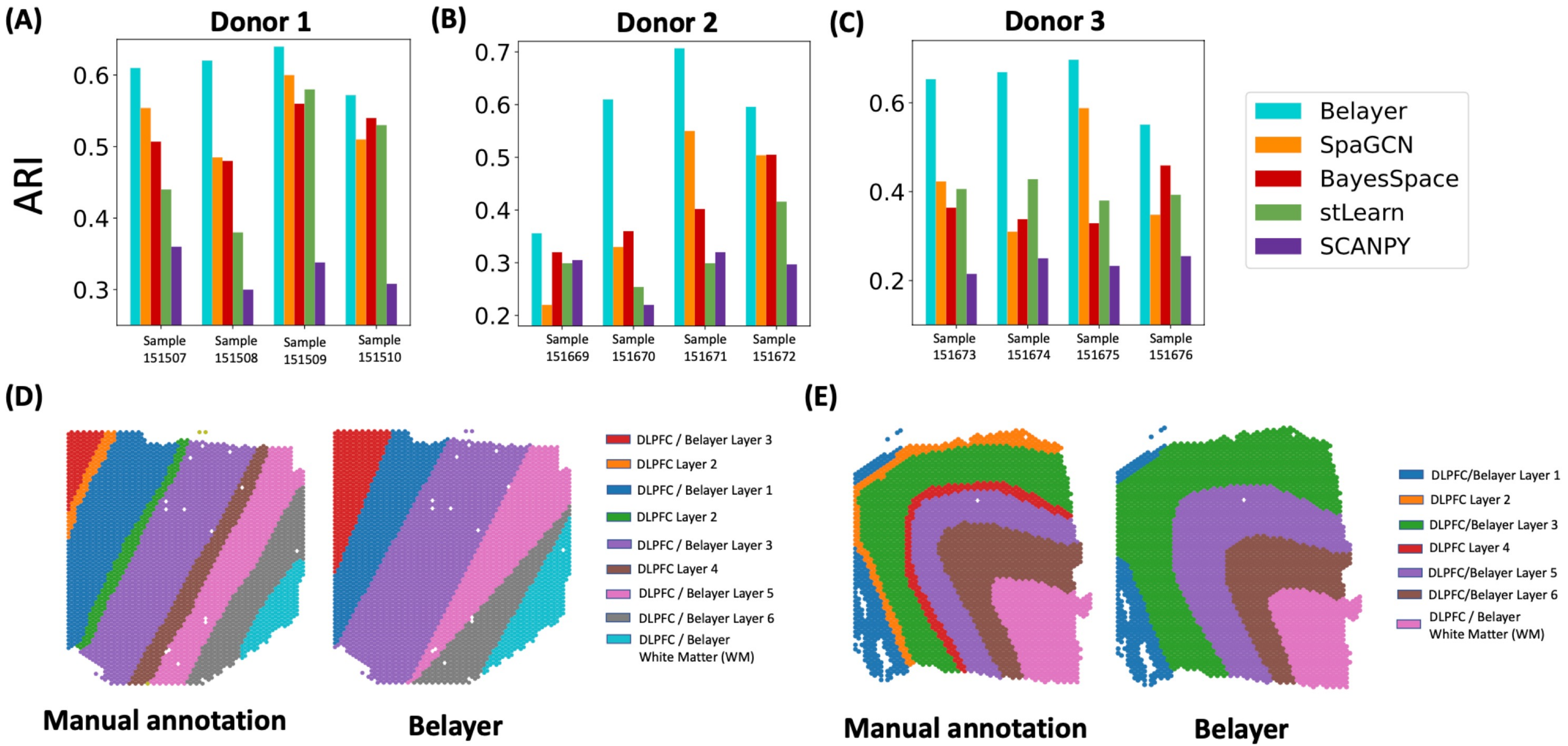
Output

Tissue layers

Spatially varying genes

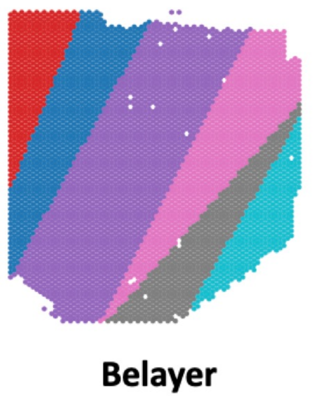
L: number of layers

Belayer accurately identifies cortical layers in human DLPFC

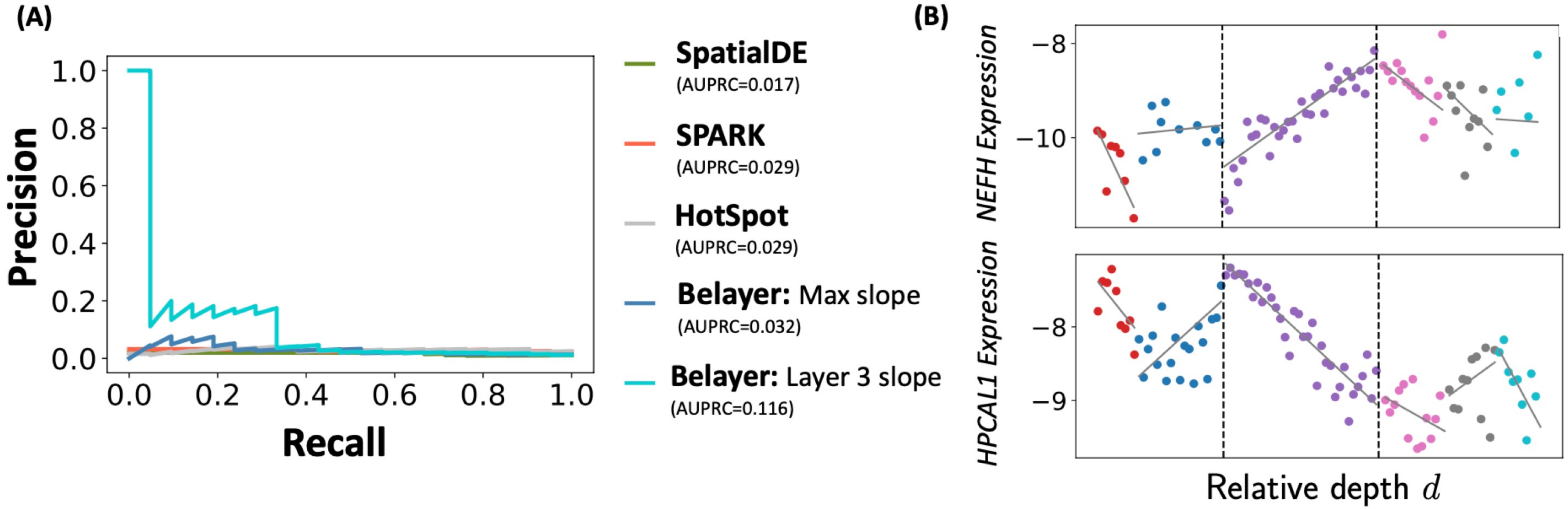


*Takeaway: Belayer (**global** model) outperforms **local** models*

Belayer identifies marker genes in human DLPFC data



We identify marker genes using slopes/discontinuities of gene expression functions f_g

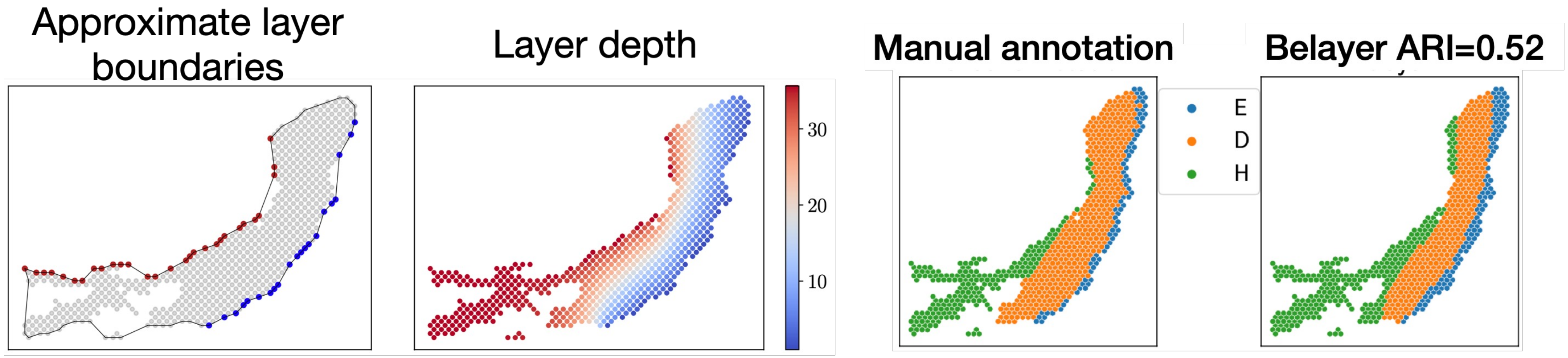


Known marker gene. Involved in neuronal damage.

Not a recorded marker gene. Involved in neuronal signaling.

*Takeaway: Belayer (**global** model) outperforms **local** models*

Mouse skin wound (10x Visium)

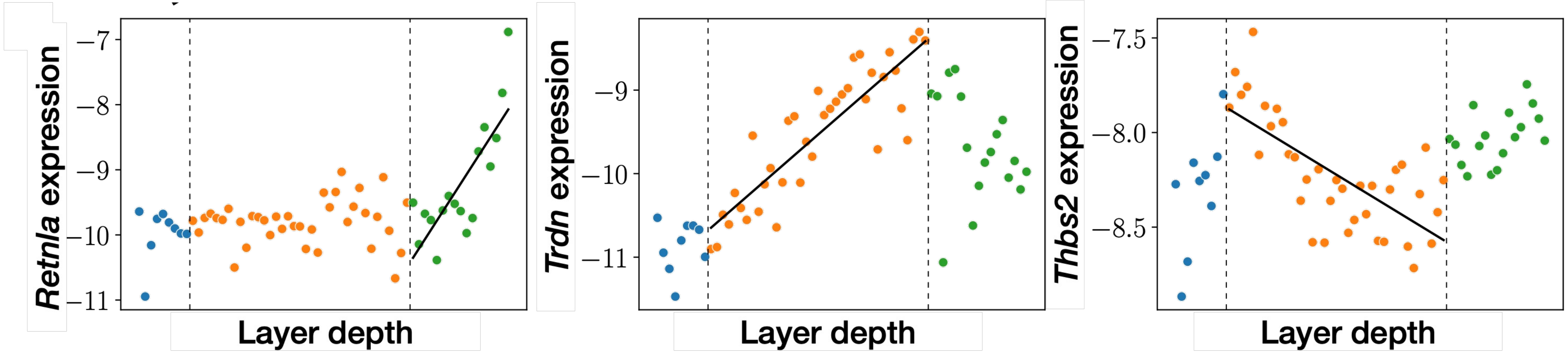


Approximate layer boundaries

Layer depth

Manual annotation

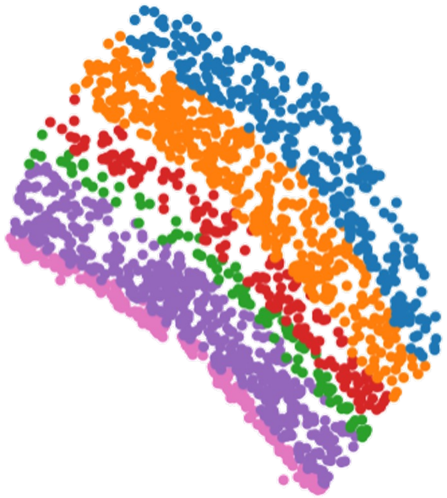
Belayer ARI=0.52



Spatially varying genes involved in muscle contraction and wound healing

Data from Foster et al. PNAS 2021

Summary of Belayer



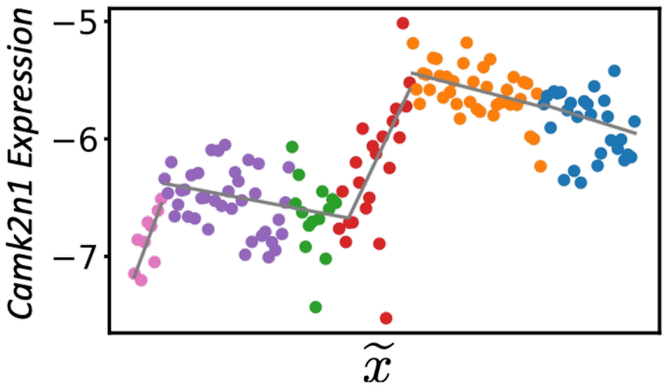
Global model of spatial gene expression for layered tissues

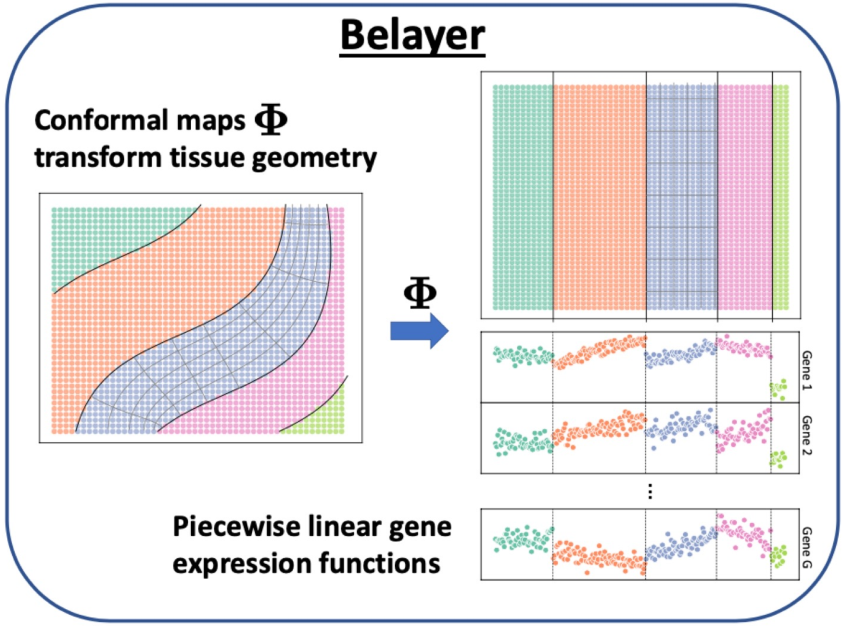
piecewise linear functions + conformal maps

Belayer simultaneously learns

- **tissue geometry** (layers) and
- **spatially varying genes** (slopes of piecewise linear functions)

from sparse SRT data





Belayer

Conformal maps Φ transform tissue geometry

Piecewise linear gene expression functions

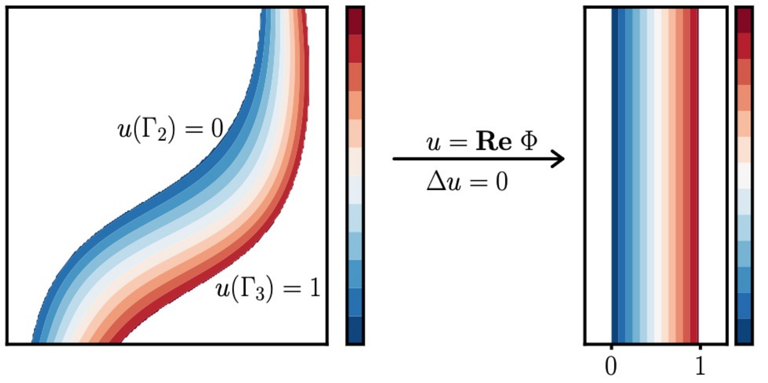
Ma*, Chitra*, et al. *Cell Systems* 2022 [Also: RECOMB 2022]

Paper

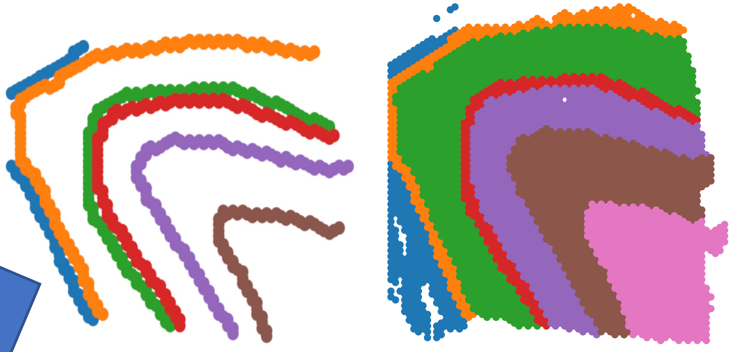
Code

Limitations of Belayer

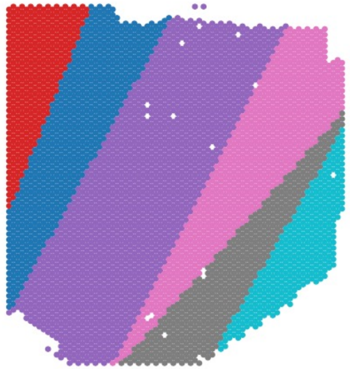
Belayer: Geometric Model



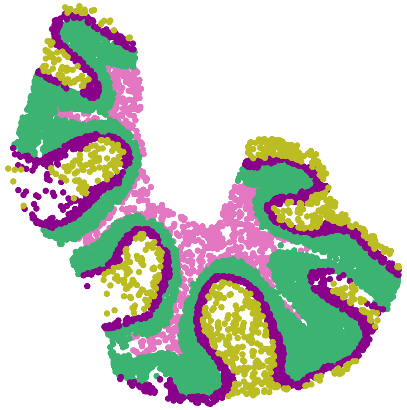
Conformal maps
Harmonic functions
Heat equation
…



✅ **Special case 1:** Manually annotated layer boundaries



✅ **Special case 2:** Layer boundaries are lines



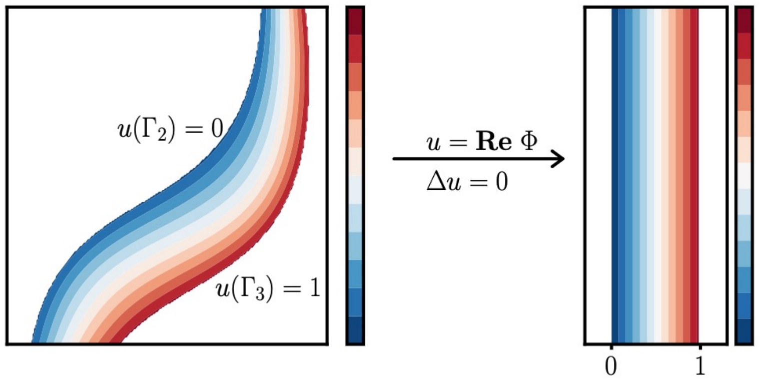
Mouse cerebellum

❌ **Cannot model**: tissues with complex or unknown boundaries

Belayer is **<u>supervised</u>** – requires *prior knowledge* of tissue geometry.

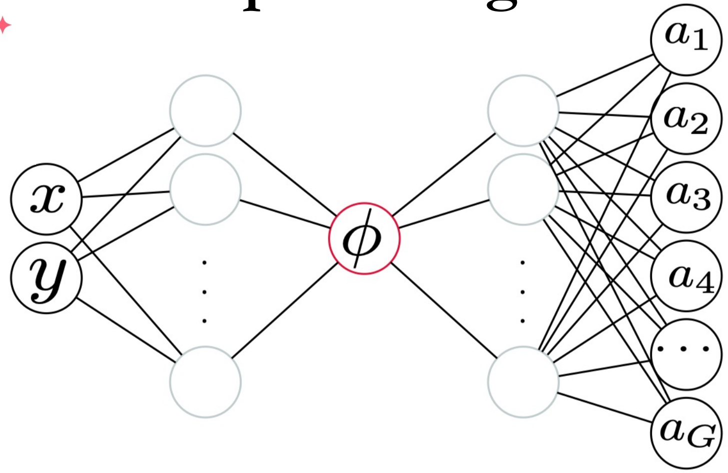
A new approach

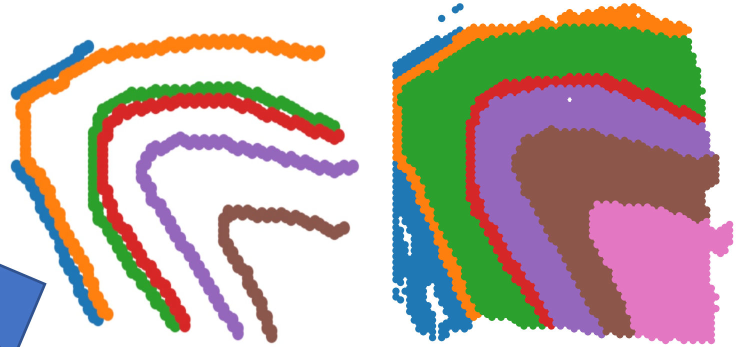
Belayer: Geometric Model



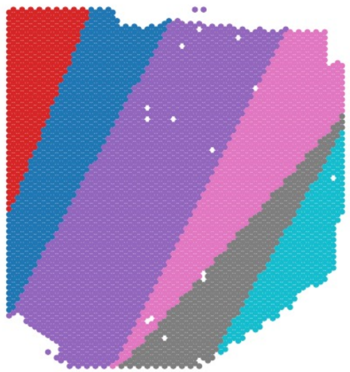
Conformal maps
Harmonic functions
Heat equation
…

Deep learning

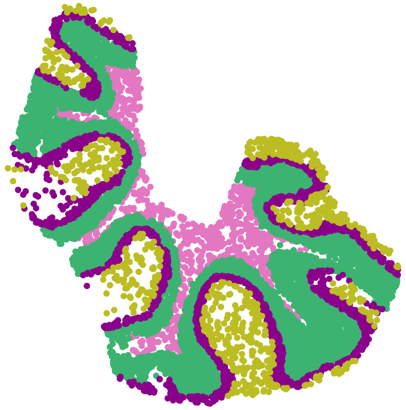




✅ **Special case 1:** Manually annotated layer boundaries



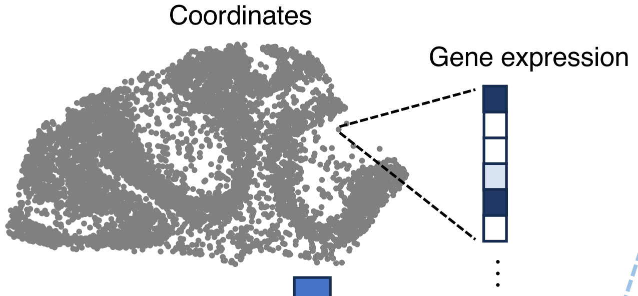
✅ **Special case 2:** Layer boundaries are lines



Mouse cerebellum

? learn layer depth without prior knowledge

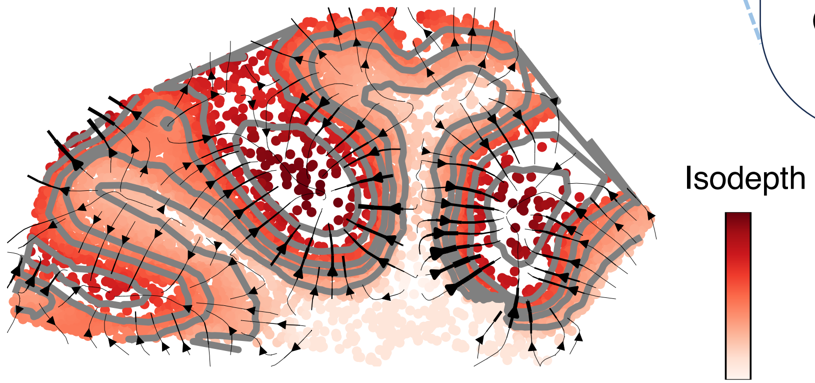
GASTON: neural network architecture



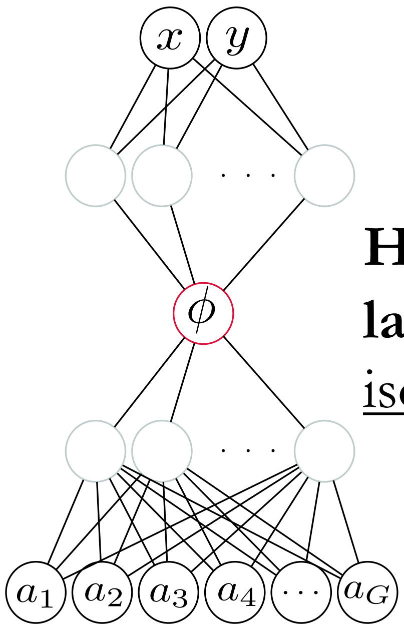
GASTON

Gradient Analysis of Spatial Transcriptomics Organization with Neural networks

Topographic map of tissue slice



Input: Spatial coordinates

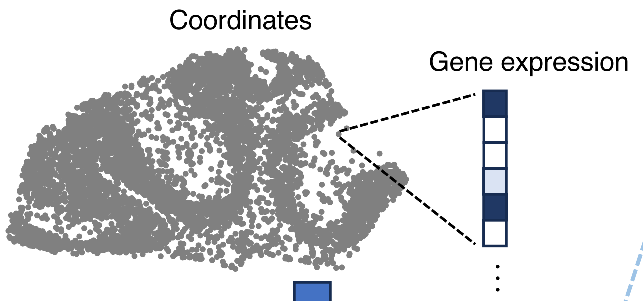


Hidden layer: isodepth

Output: Gene expression

Training is **unsupervised**!
(like an auto-encoder)

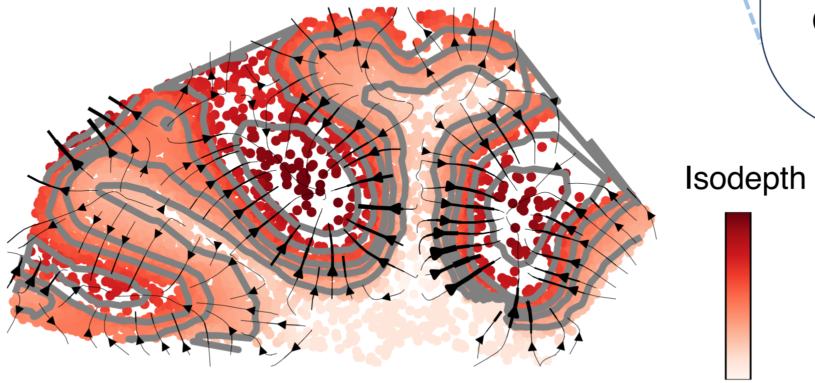
Isodepth defines "topography" of gene expression

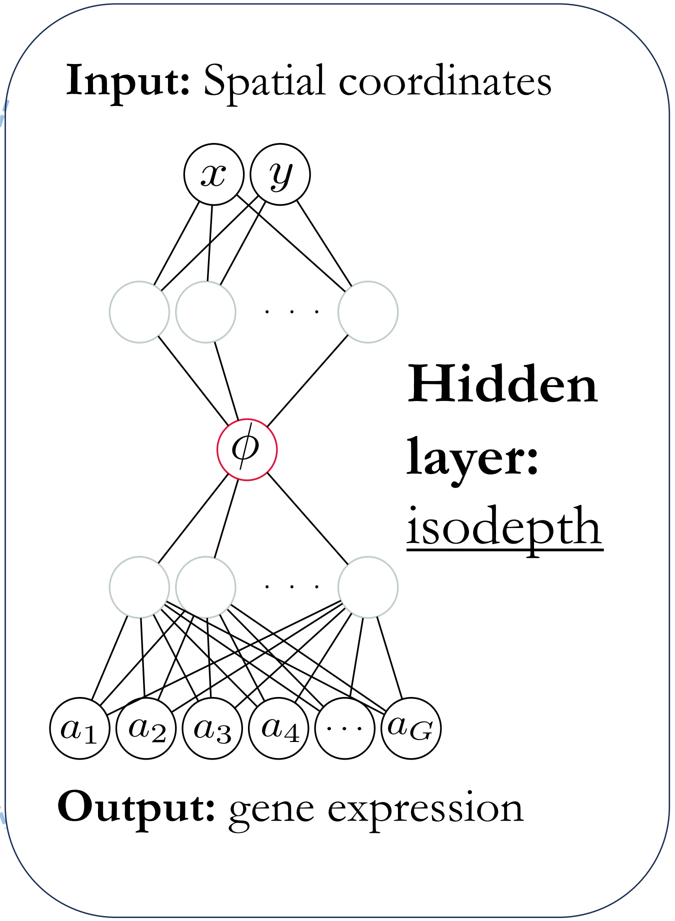


GASTON

Gradient Analysis of Spatial Transcriptomics Organization with Neural networks

Topographic map of tissue slice



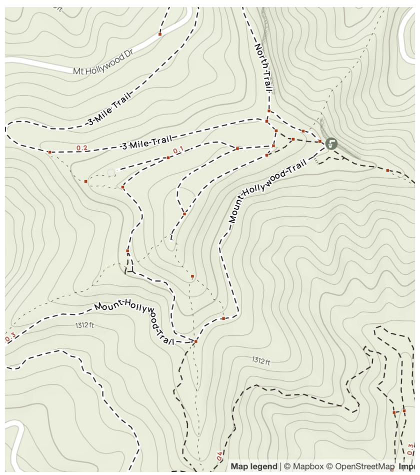


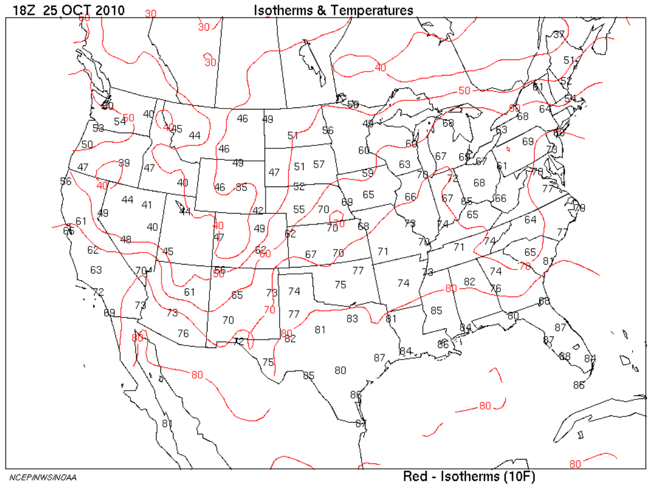
Input: Spatial coordinates

Hidden layer: isodepth

Output: gene expression

Training is **unsupervised**!
(like an auto-encoder)





Isodepth = contours of equal depth: $\phi = c$

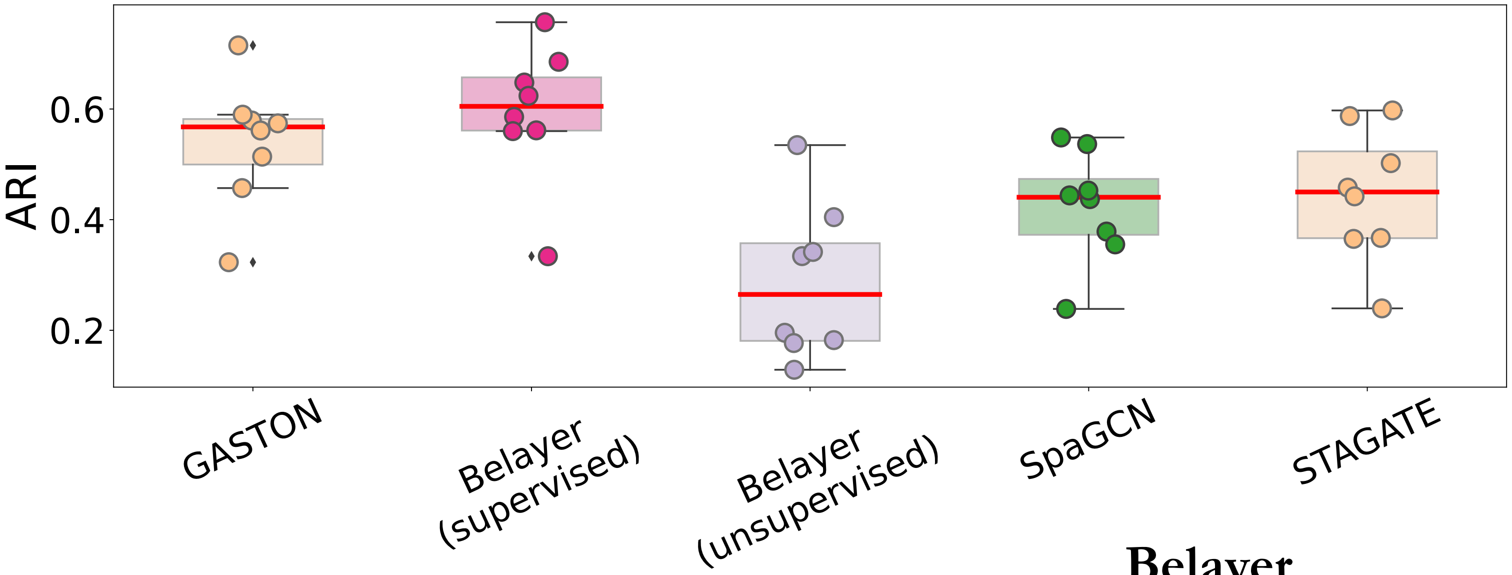
- Generalizes *relative depth* from Belayer
- *Neural field* model (used in computer vision/graphics)

Spatial gradients $\nabla\phi$ (gradient of isodepth)

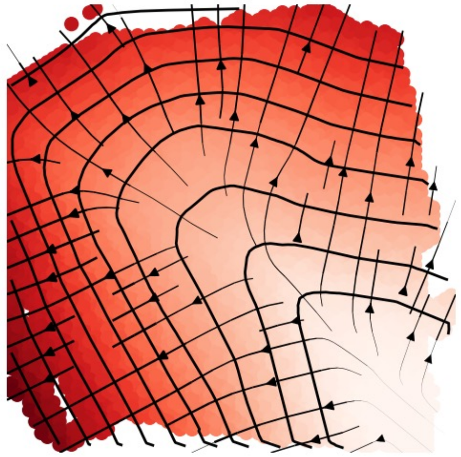
- Directions of maximum change in gene expression
- Gradient field $\nabla\phi$ is "*conservative*" (no curl)

Gene expression functions $f_g(\phi(x,y))$

Human DLPFC: GASTON outperforms other neural networks and unsupervised Belayer

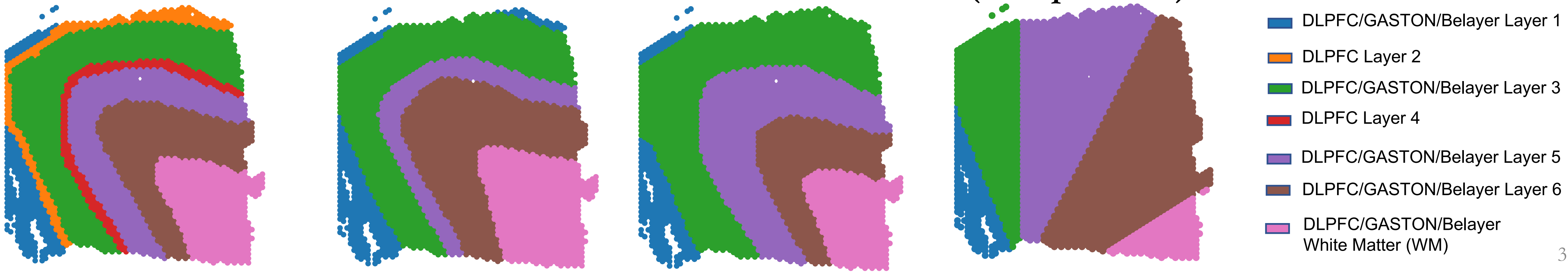


GASTON isodepth



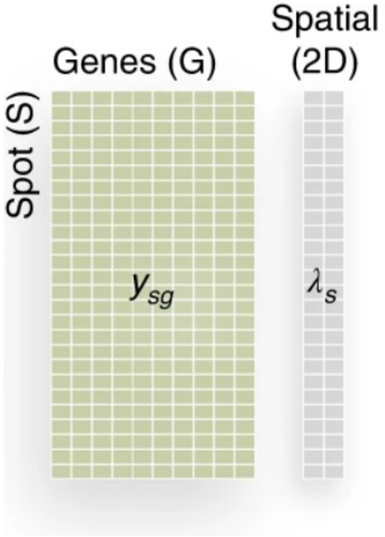
0.99 correlation with (supervised) Belayer depth!

Manual annotation **GASTON** **Belayer (supervised)** **Belayer (unsupervised)**



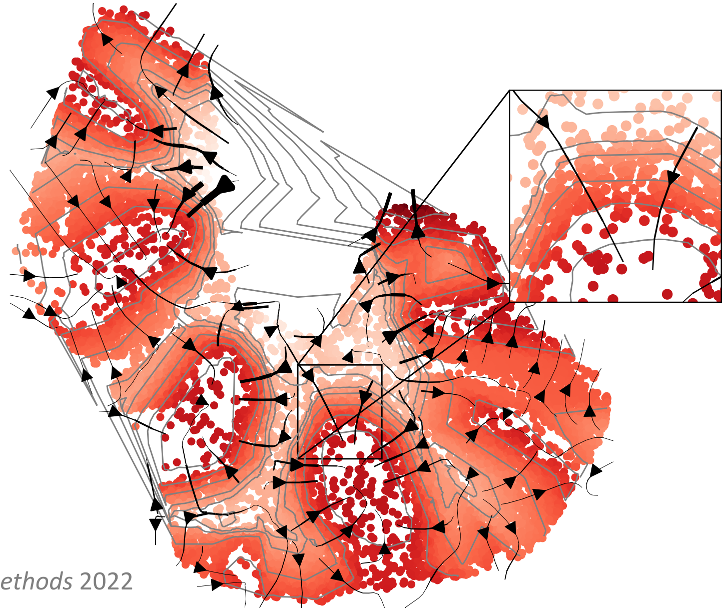
DLPFC/GASTON/Belayer Layer 1
DLPFC Layer 2
DLPFC/GASTON/Belayer Layer 3
DLPFC Layer 4
DLPFC/GASTON/Belayer Layer 5
DLPFC/GASTON/Belayer Layer 6
DLPFC/GASTON/Belayer White Matter (WM)

GASTON: Mouse Cerebellum (Slide-seqV2)

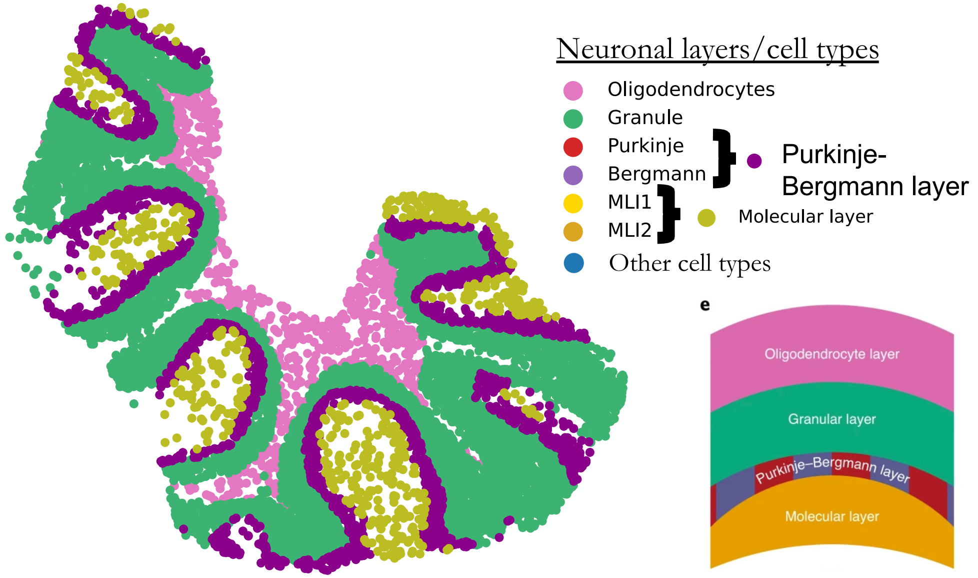


GASTON

Topographic map



Spatial domains

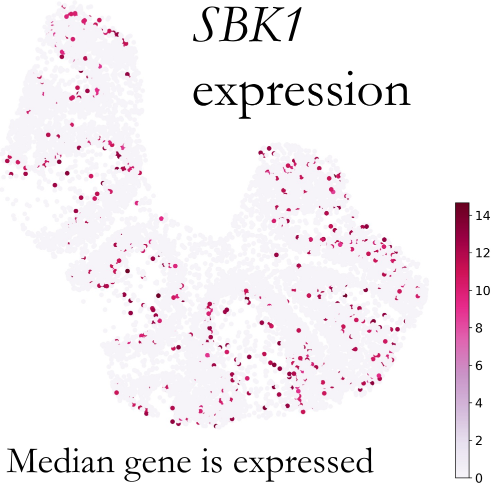


23,096 genes × 9,985 spots

Cable et al., *Nature Methods* 2022

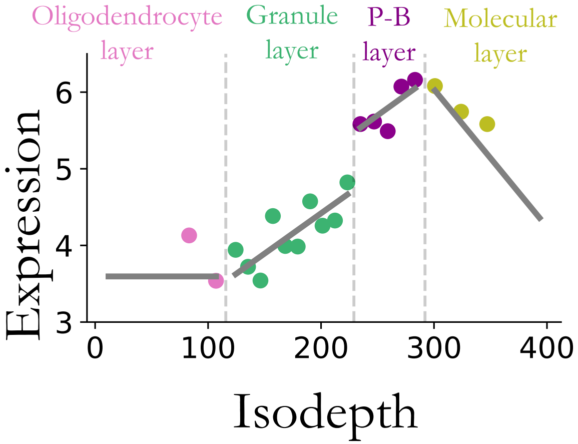
Baizer, Front. Hum. Neurosci. 2014

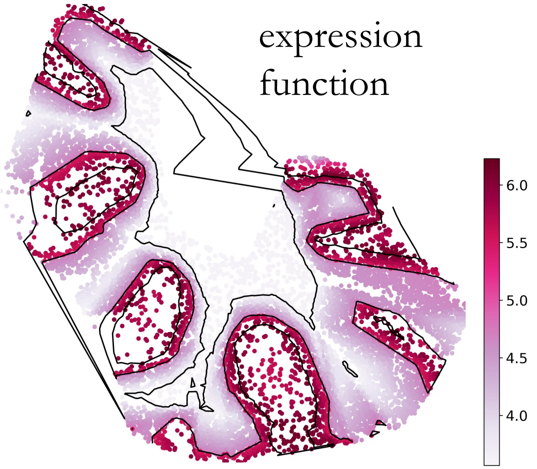
Median spot has 370 UMIs



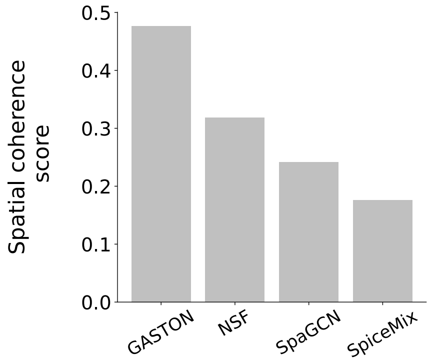
SBK1 expression

Median gene is expressed in 0.2% of spots



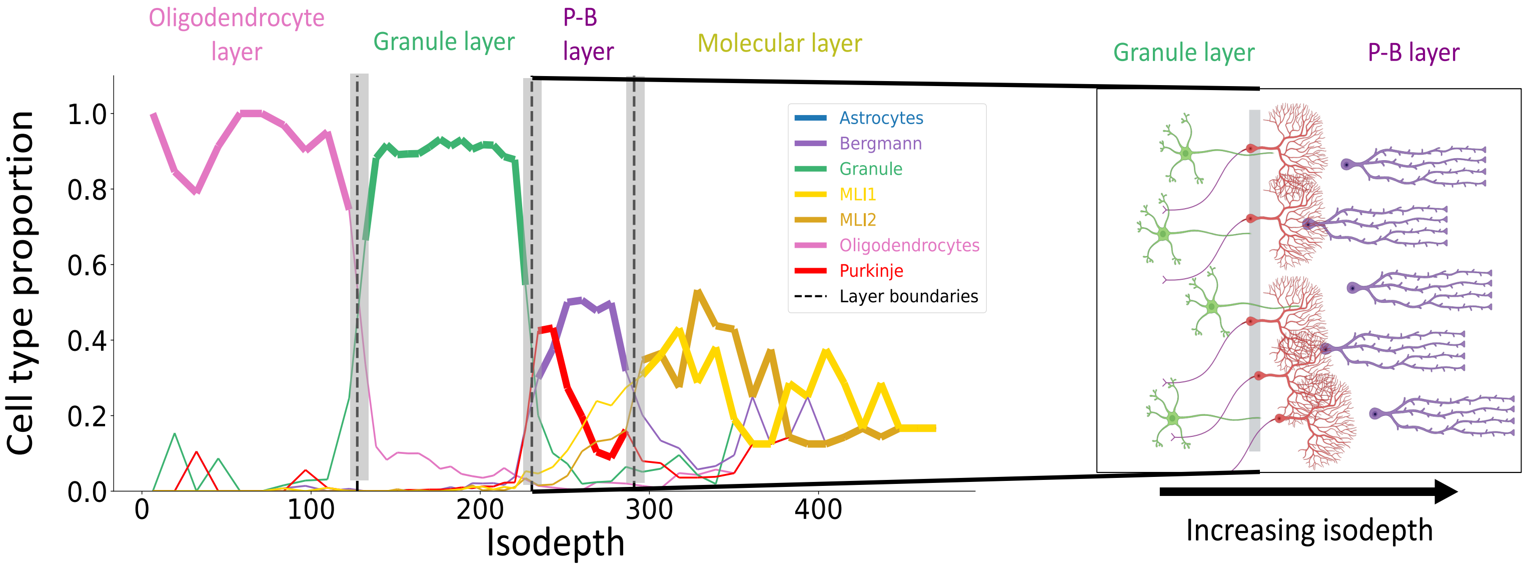


GASTON *SBK1* expression function

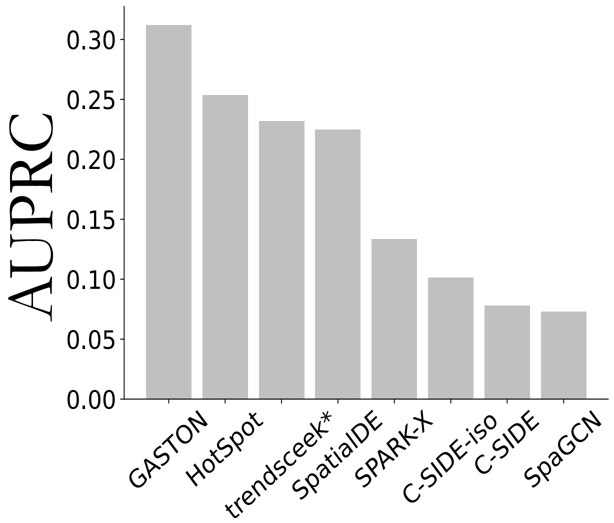


Cell type and gene expression gradients

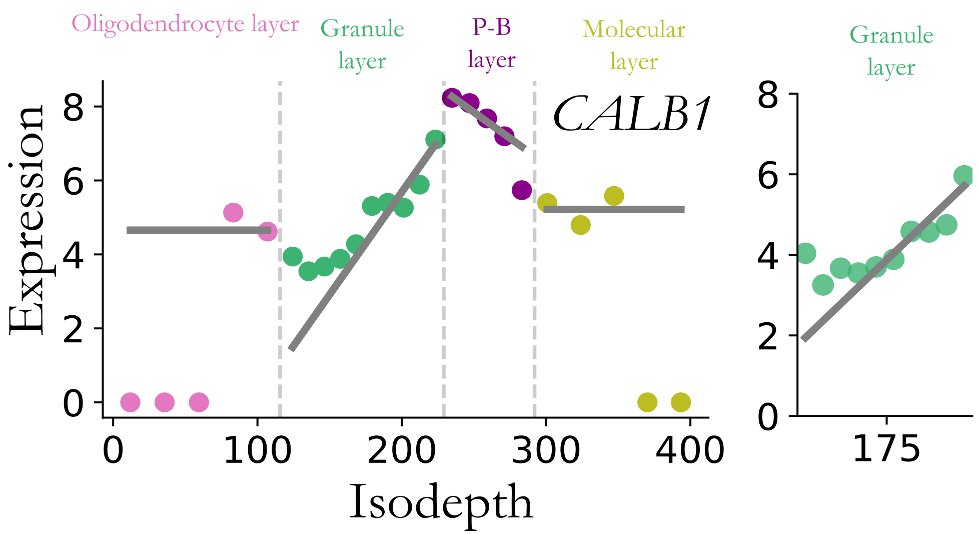
GASTON



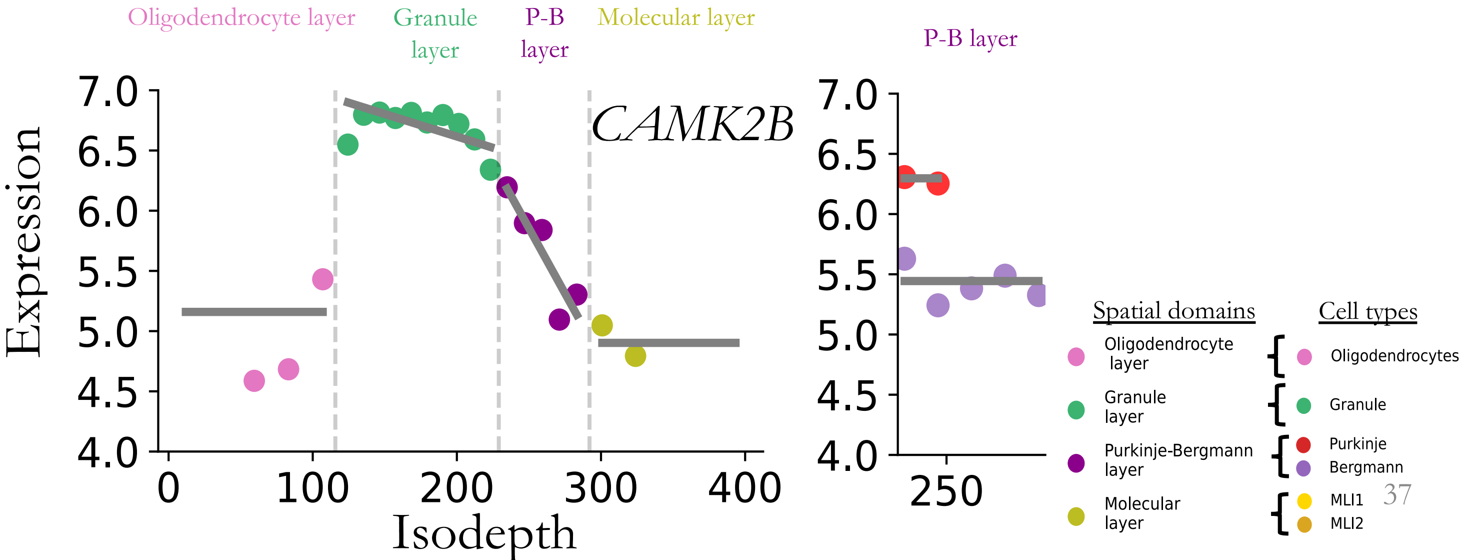
Marker gene identification



Cell type-attributable gradient



Other-attributable gradient

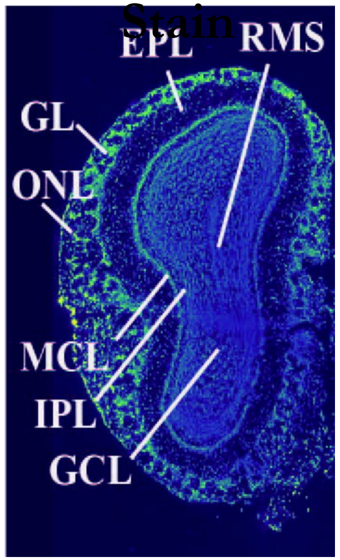


Olfactory bulb (Stereo-seq)

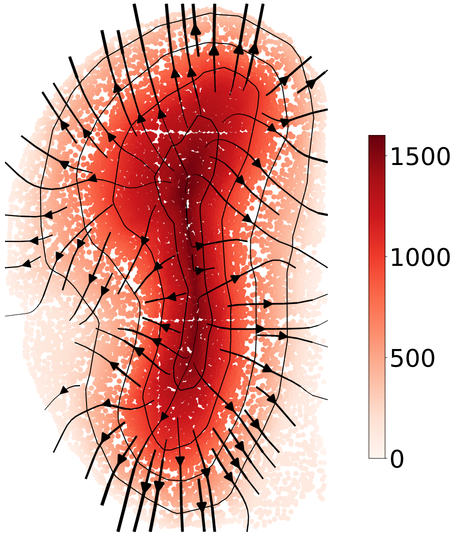
9,825 spots × 27,106 genes

- ● Olfactory nerve layer (ONL)
- ● Glomerular layer (GL)
- ● External plexiform layer (EPL)
- ● Mitral cell layer (MCL)
- ● Internal plexiform layer (IPL)
- ● Granule cell layer (GCL)
- ● Rostral migratory stream (RMS)

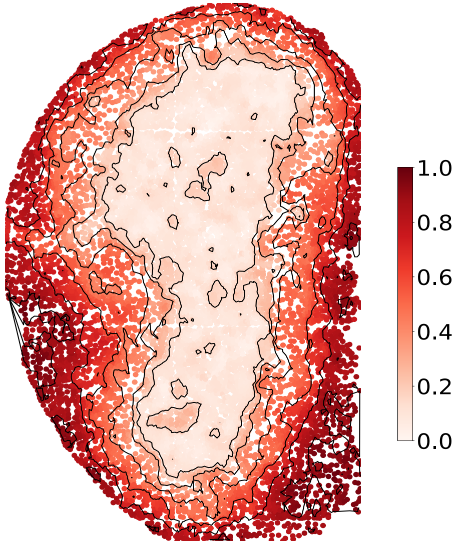
DAPI Stain



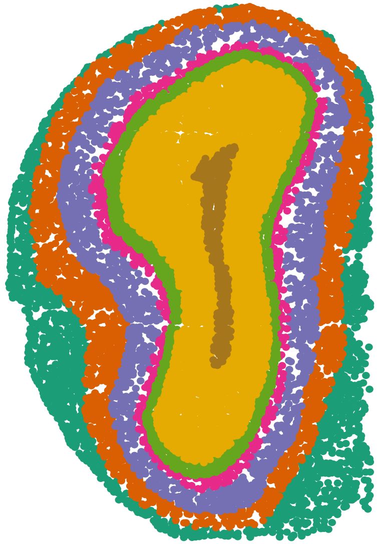
Isodepth and (negative) spatial gradients



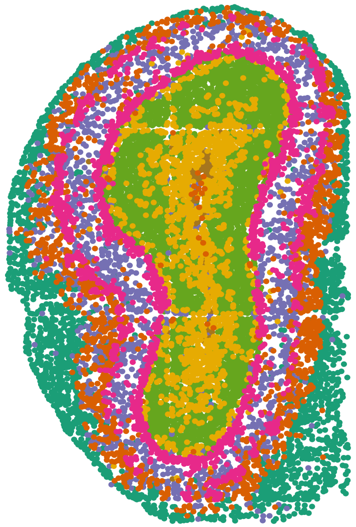
SpaceFlow
(diffusion pseudotime)

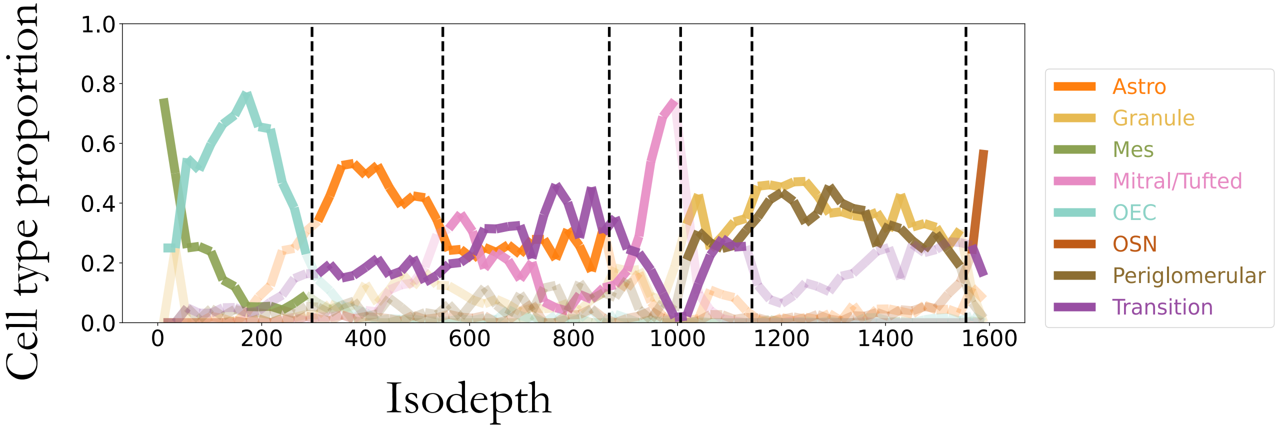


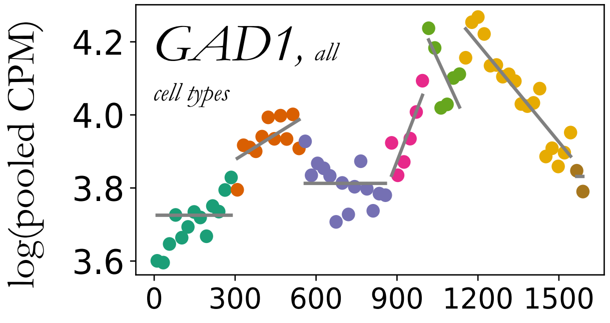
GASTON



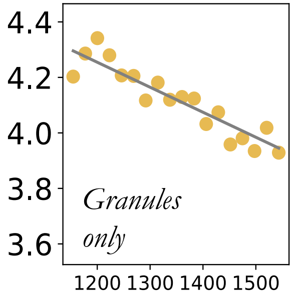
SpaGCN



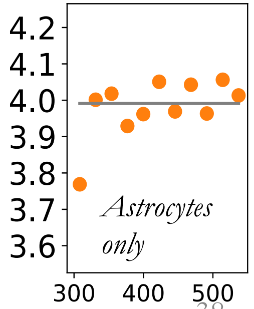




Cell type-*attributable* gradient



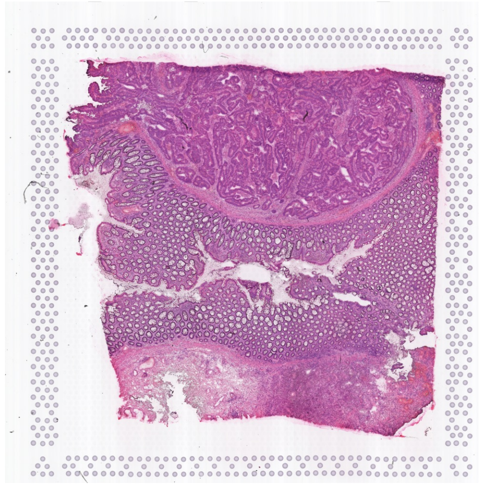
Other attributable gradient

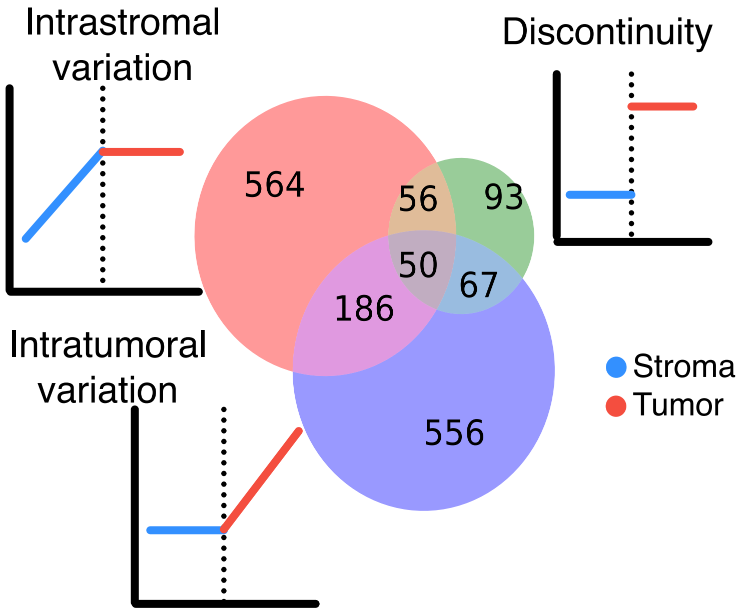


Isodepth

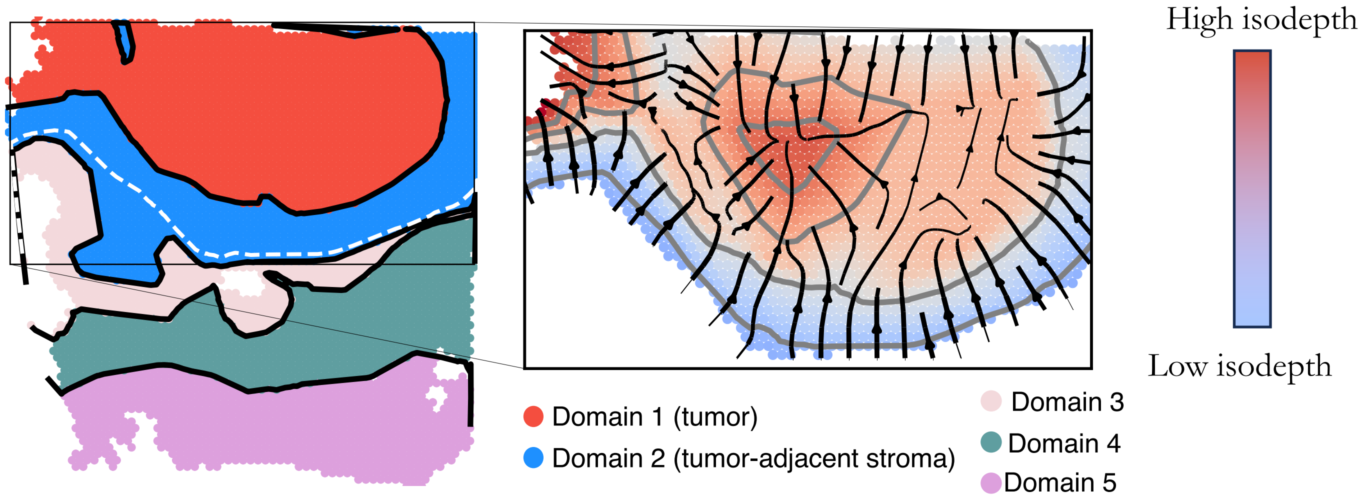
GASTON identifies gradients in tumor microenvironment

Colorectal tumor slice (stage IV)
(Wu et al, Cancer Discovery 2022)

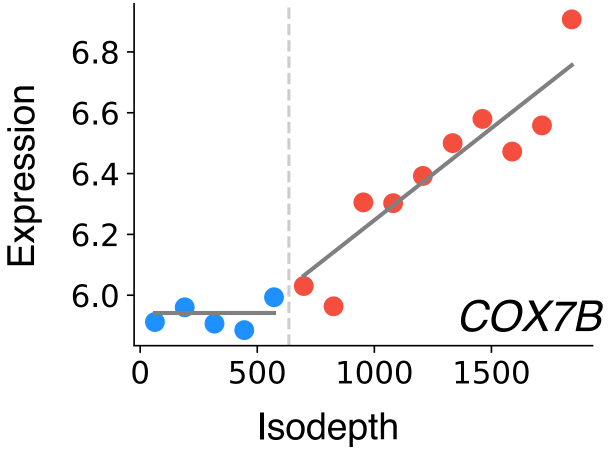




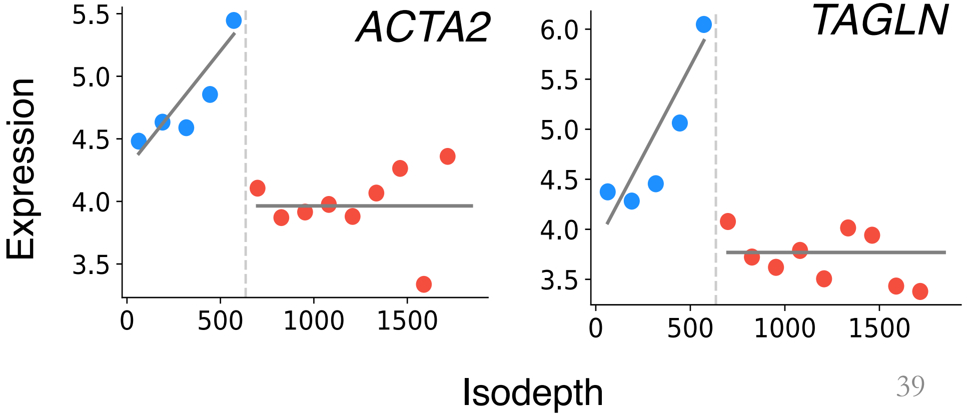
GASTON: spatial domains + isodepth



Intratumoral variation
aerobic metabolism



Discontinuity/Intrastromal variation-
epithelial-mesenchymal transition (EMT)



Summary: **GASTON**

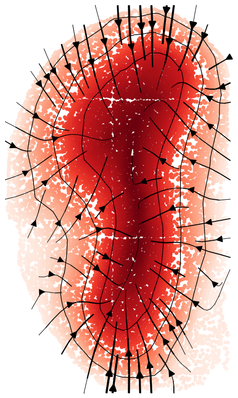
- Isodepth describes **topographic map** and **spatial gradients** of gene expression within tissue slice

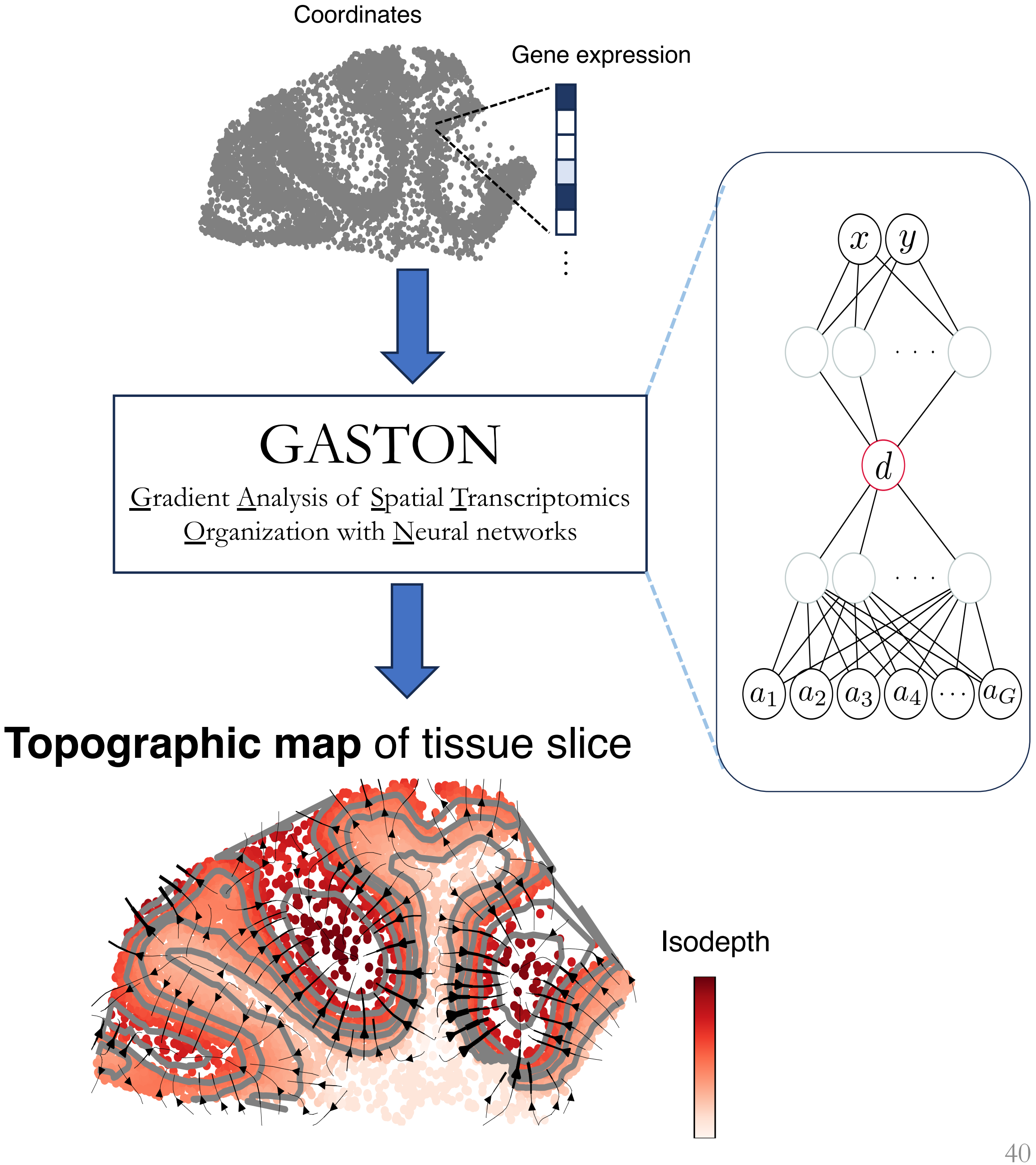
- GASTON: **unsupervised** deep learning algorithm to learn isodepth
  - Uncovers spatial domains and gradients of gene expression/cell type

Chitra et al. In review at *Nature Methods*
[Also: RECOMB 2024]

Preprint

Code





Coordinates

Gene expression

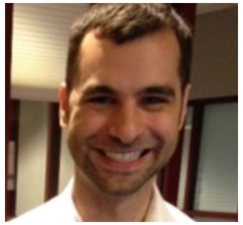
GASTON

Gradient Analysis of Spatial Transcriptomics Organization with Neural networks

Topographic map of tissue slice

Isodepth

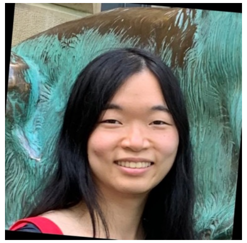
Identifying altered subnetworks (network anomalies)



Matt Reyna

Rebecca Elyanow

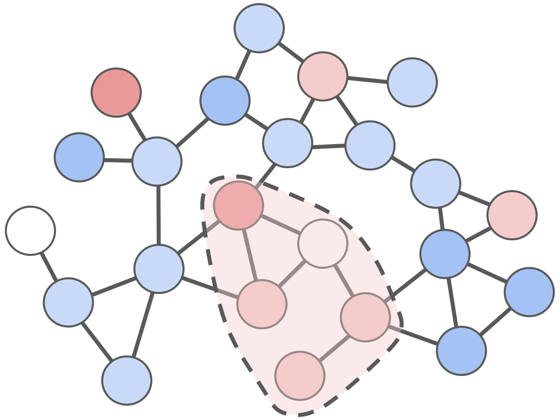
Tyler Park



Kimberly Ding

Jasper C. H. Lee

Ben Raphael



Reyna*, **Chitra***, Elyanow, Raphael. *RECOMB 2020 + Journal of Computational Biology.*

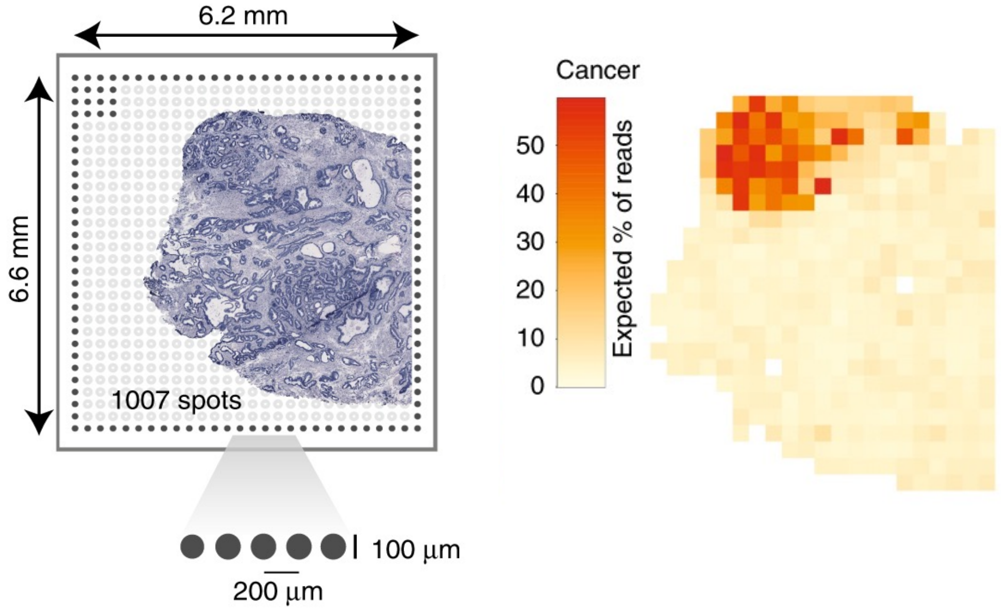
Chitra, Ding, Lee, Raphael. *ICML 2021.*

Chitra*, Park*, Raphael. *RECOMB 2022 + Journal of Computational Biology.*

* indicates joint first authorship

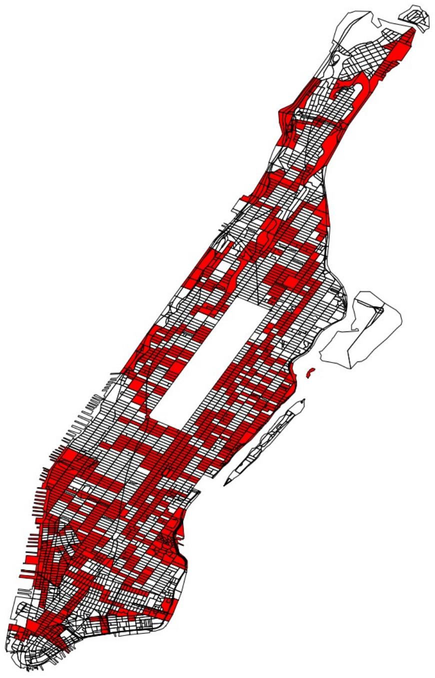
Spatial anomaly detection

Spatial anomaly in biology:



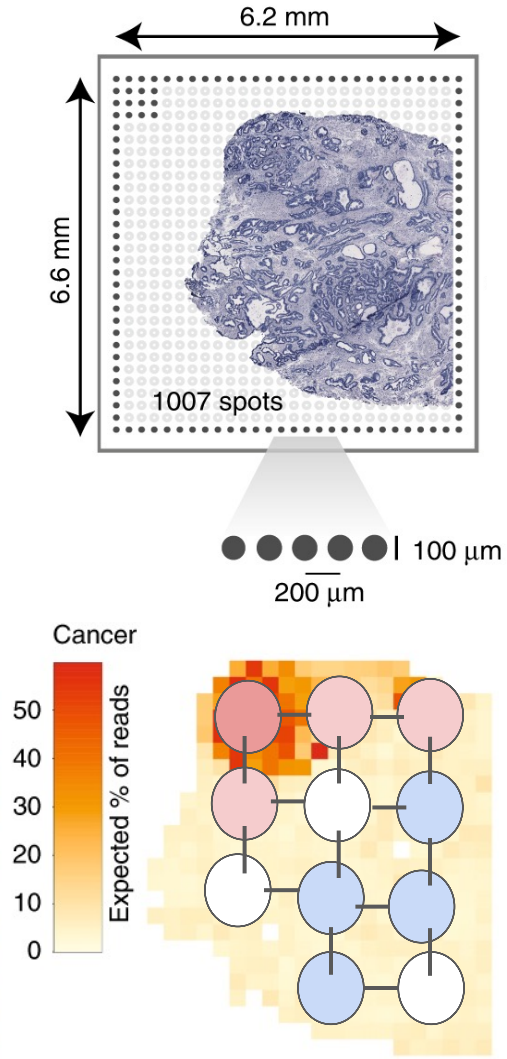
Tumor detection (prostate cancer)

Spatial anomaly in epidemiology:



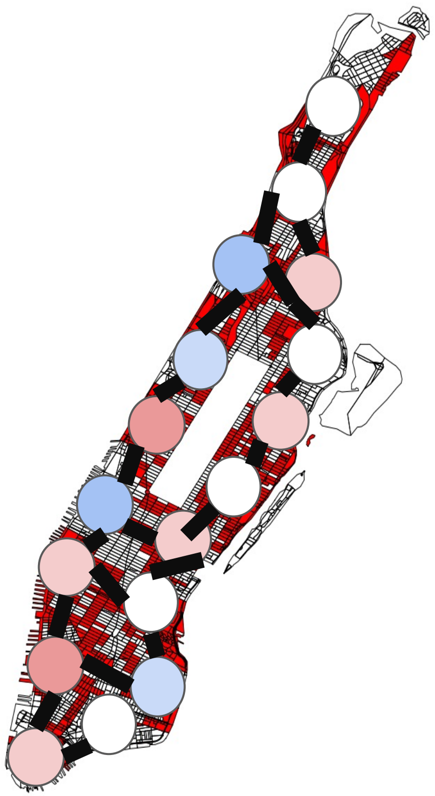
Disease hotspots identification (breast cancer incidence, NYC)

Network anomaly detection



Tumor detection

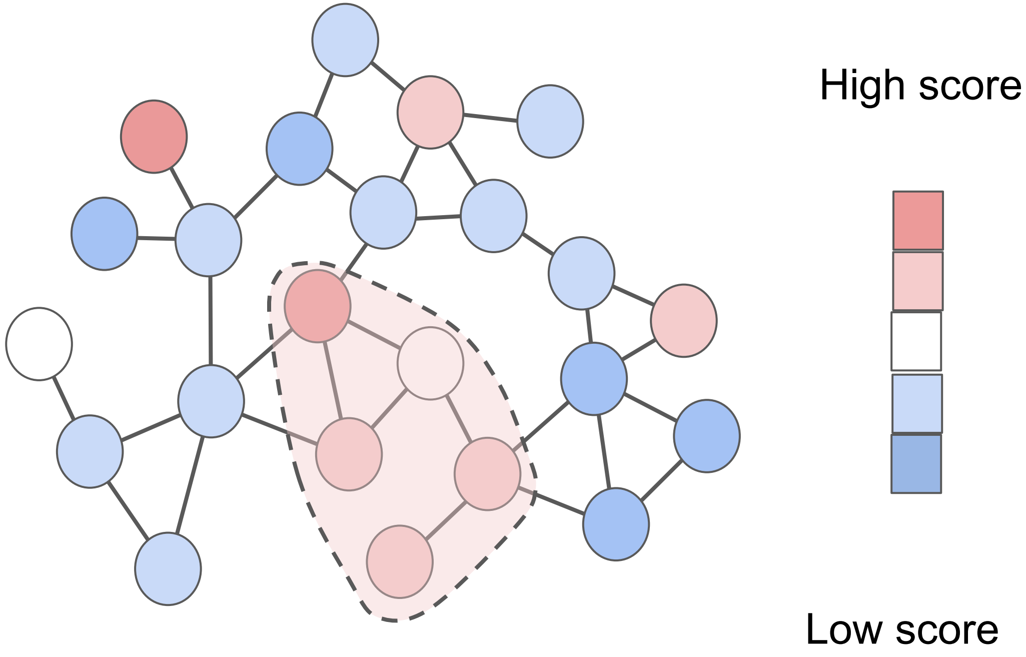
(prostate cancer)



Epidemiology:

Disease hotspots

(breast cancer, NYC)



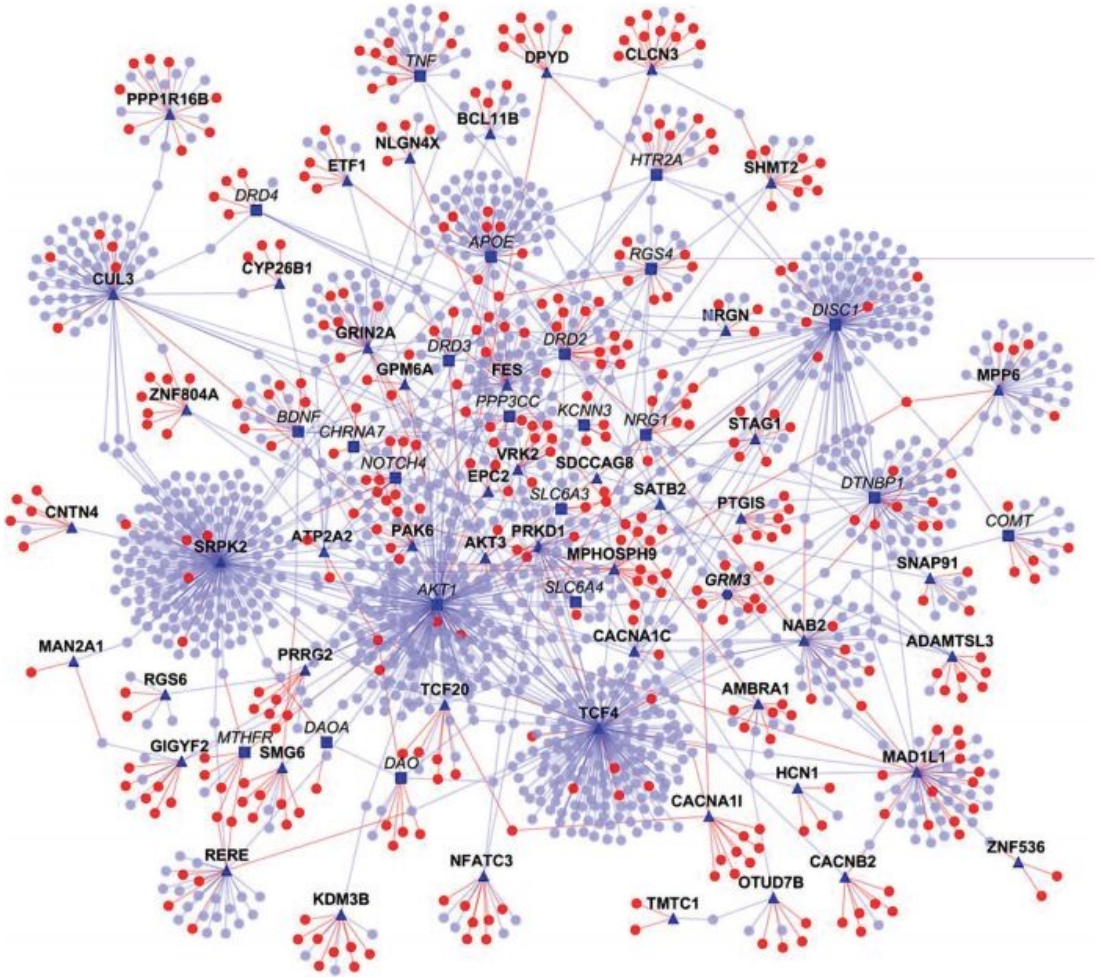
Network anomaly detection

- Vertices = spatial locations (e.g. tissue spots, census tracts)
    - Edges connect spatially adjacent vertices
- Score: % cancer cells OR disease incidence OR …

<u>Anomalies</u>: subnetworks with large score

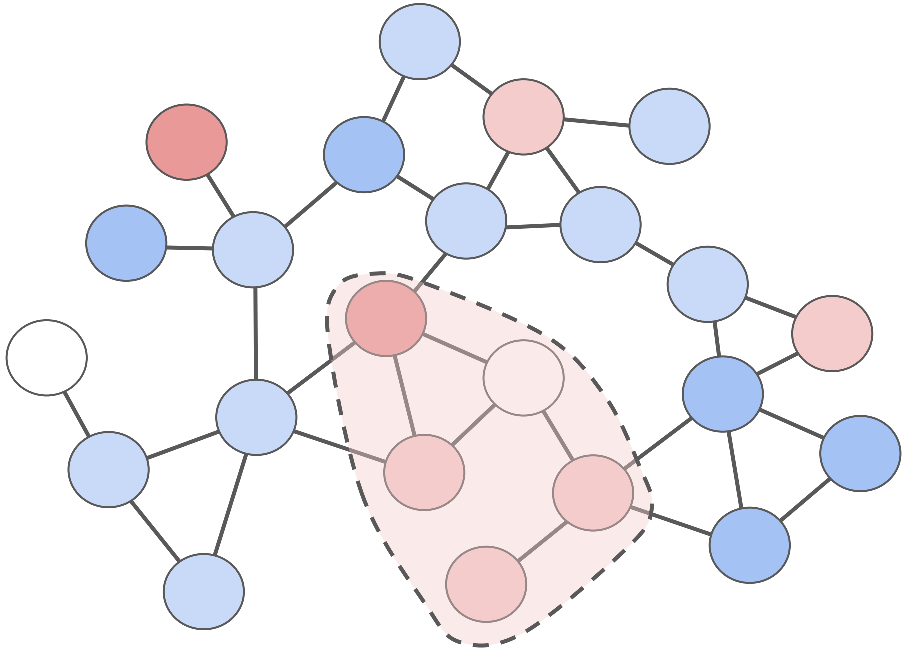
Protein-Protein Interaction (PPI) Networks

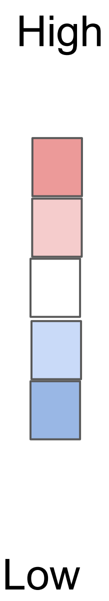
Vertices: proteins
Edges: physical interactions between proteins



Altered Subnetwork Problem (ASP)

(also called network anomalies, network modules, active subnetworks,)



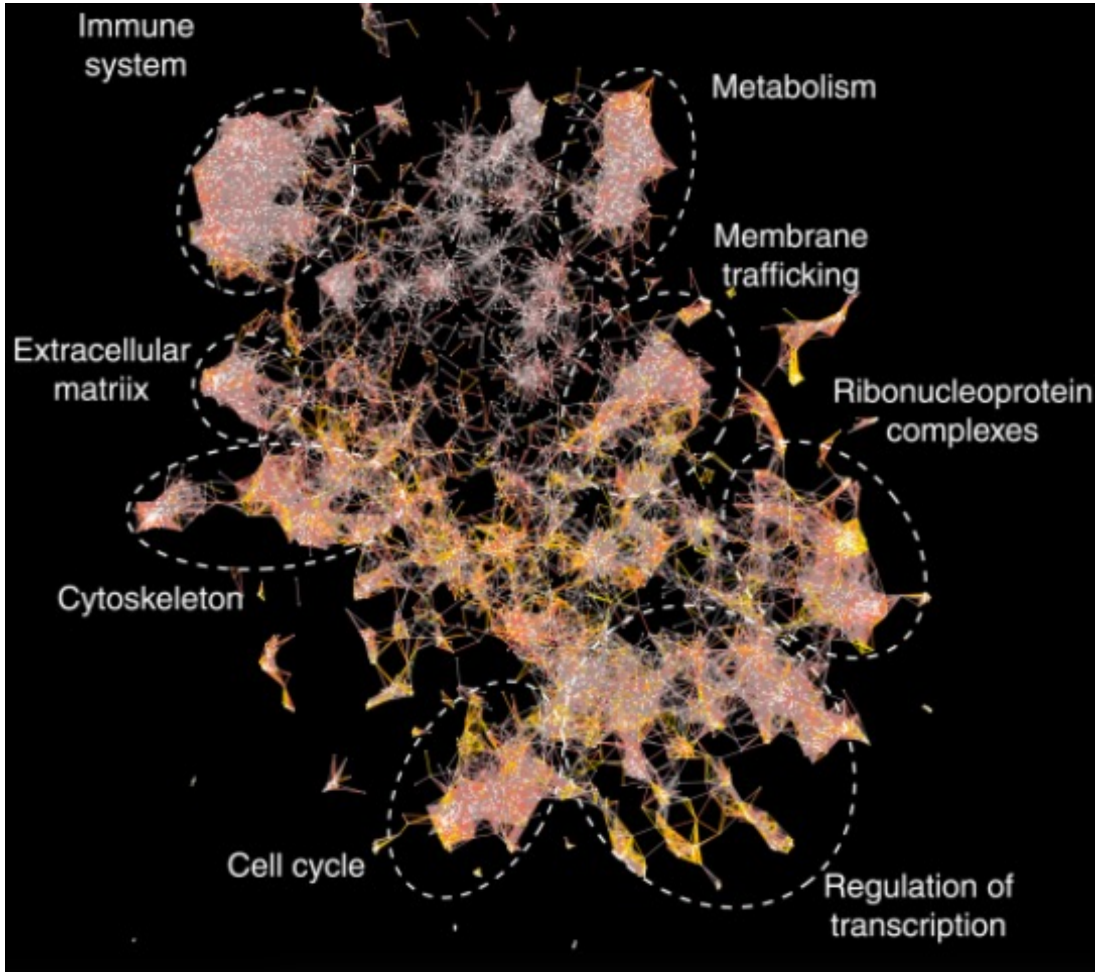


Given:

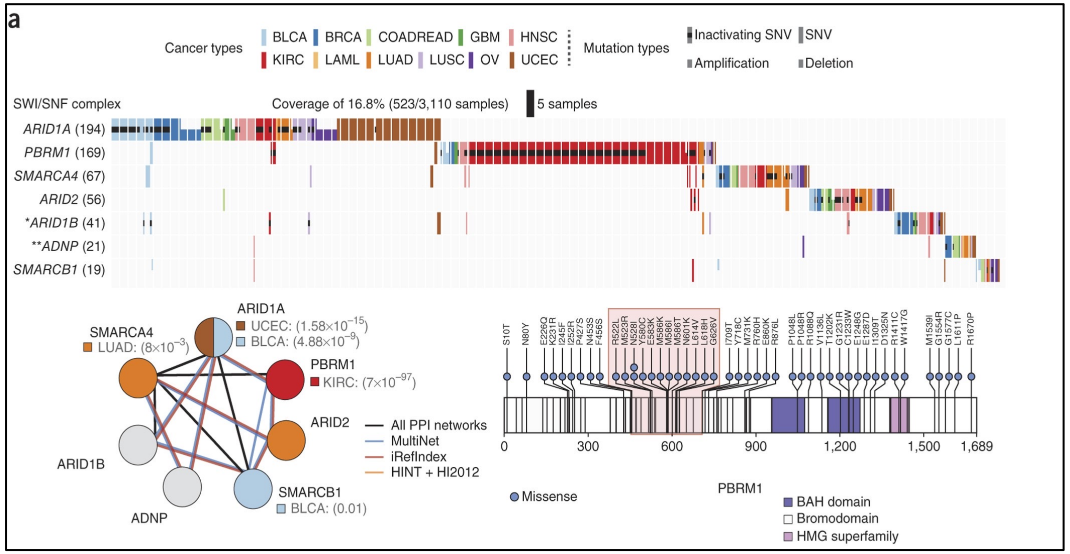
1) Interaction network G = (V,E)
2) Vertex scores X_v
   - eg from mutations, differential expression, …

Goal: Identify **high-scoring subnetworks** of G ("altered subnetworks")

Altered subnetworks reveal interacting genes relevant to complex traits+diseases

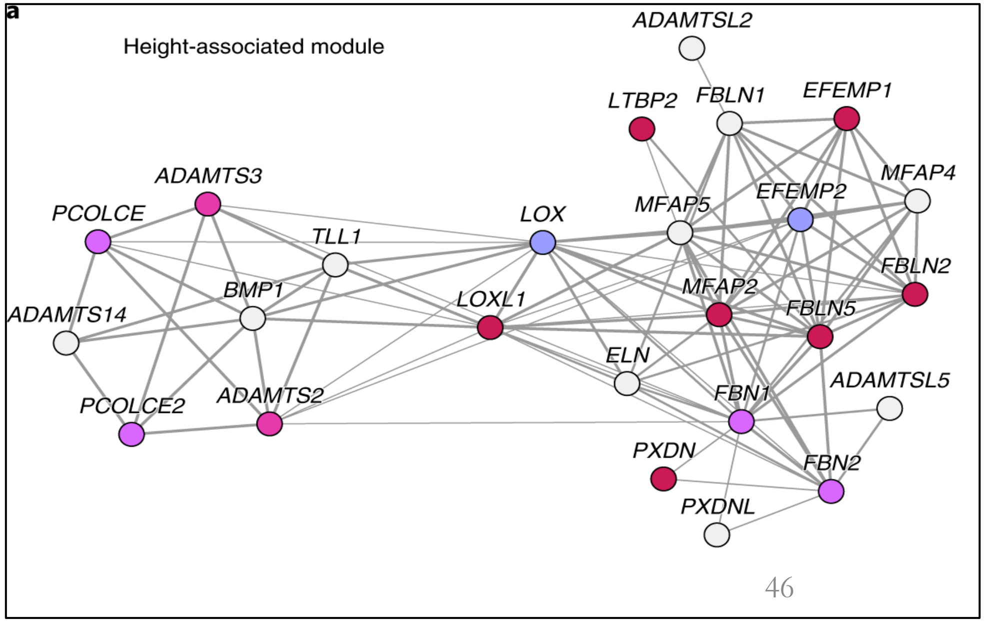


Somatic mutations in cancer



Leiserson, Vandin et al (Nature Genetics 2015)

Complex traits (e.g. height, diabetes, …)



Choobdar, Ahsen, Crawford et al (Nature Methods 2019)

Many algorithms developed over past 20 years for identifying altered subnetworks

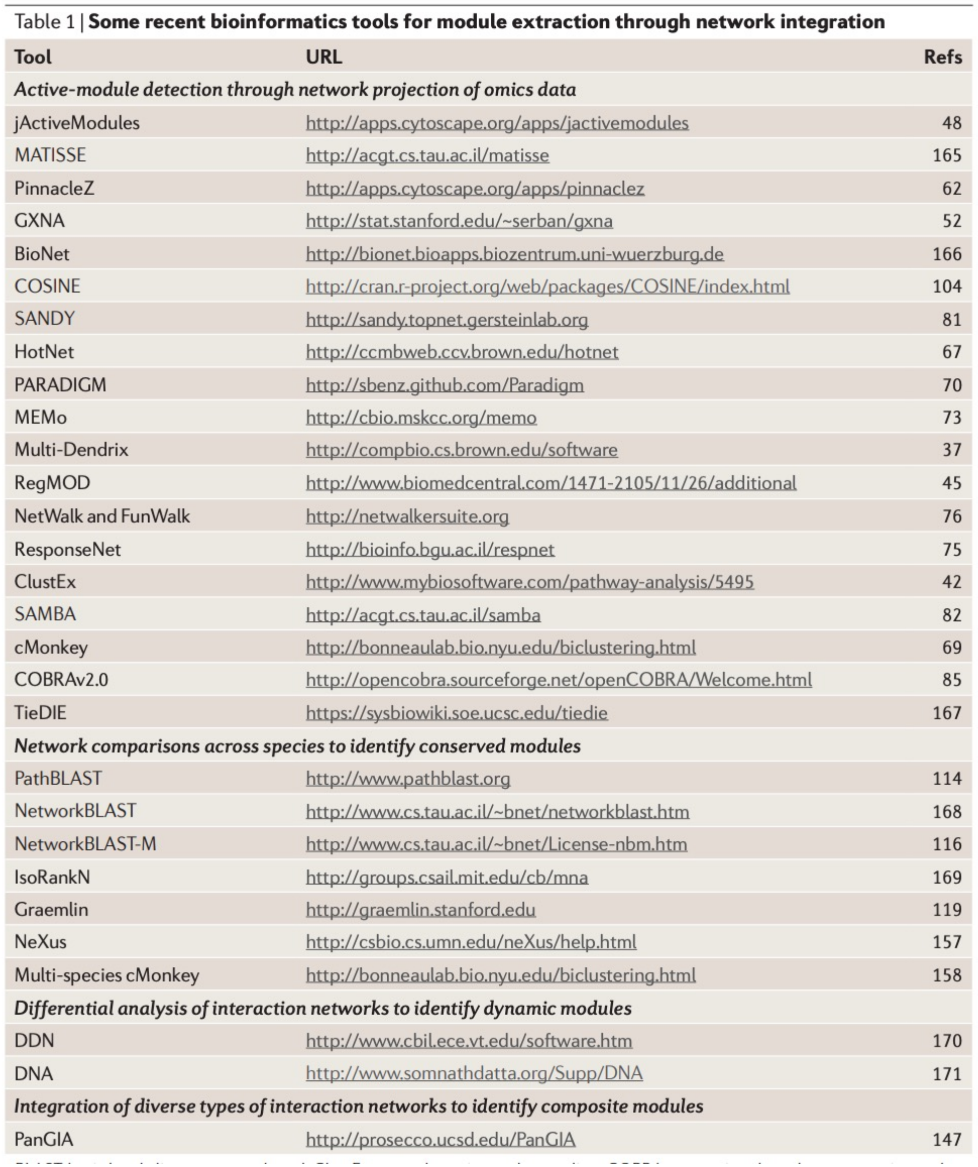


Table 1 | **Some recent bioinformatics tools for module extraction through network integration**

Tool	URL	Refs
Active-module detection through network projection of omics data		
jActiveModules	http://apps.cytoscape.org/apps/jactivemodules	48
MATISSE	http://acgt.cs.tau.ac.il/matisse	165
PinnacleZ	http://apps.cytoscape.org/apps/pinnaclez	62
GXNA	http://stat.stanford.edu/~serban/gxna	52
BioNet	http://bionet.bioapps.biozentrum.uni-wuerzburg.de	166
COSINE	http://cran.r-project.org/web/packages/COSINE/index.html	104
SANDY	http://sandy.topnet.gersteinlab.org	81
HotNet	http://ccmbweb.ccv.brown.edu/hotnet	67
PARADIGM	http://sbenz.github.com/Paradigm	70
MEMo	http://cbio.mskcc.org/memo	73
Multi-Dendrix	http://compbio.cs.brown.edu/software	37
RegMOD	http://www.biomedcentral.com/1471-2105/11/26/additional	45
NetWalk and FunWalk	http://netwalkersuite.org	76
ResponseNet	http://bioinfo.bgu.ac.il/respnet	75
ClustEx	http://www.mybiosoftware.com/pathway-analysis/5495	42
SAMBA	http://acgt.cs.tau.ac.il/samba	82
cMonkey	http://bonneaulab.bio.nyu.edu/biclustering.html	69
COBRAv2.0	http://opencobra.sourceforge.net/openCOBRA/Welcome.html	85
TieDIE	https://sysbiowiki.soe.ucsc.edu/tiedie	167
Network comparisons across species to identify conserved modules		
PathBLAST	http://www.pathblast.org	114
NetworkBLAST	http://www.cs.tau.ac.il/~bnet/networkblast.htm	168
NetworkBLAST-M	http://www.cs.tau.ac.il/~bnet/License-nbm.htm	116
IsoRankN	http://groups.csail.mit.edu/cb/mna	169
Graemlin	http://graemlin.stanford.edu	119
NeXus	http://csbio.cs.umn.edu/neXus/help.html	157
Multi-species cMonkey	http://bonneaulab.bio.nyu.edu/biclustering.html	158
Differential analysis of interaction networks to identify dynamic modules		
DDN	http://www.cbil.ece.vt.edu/software.htm	170
DNA	http://www.somnathdatta.org/Supp/DNA	171
Integration of diverse types of interaction networks to identify composite modules		
PanGIA	http://prosecco.ucsd.edu/PanGIA	147

Mitra *et al*, Nature Reviews Genetics (2013)

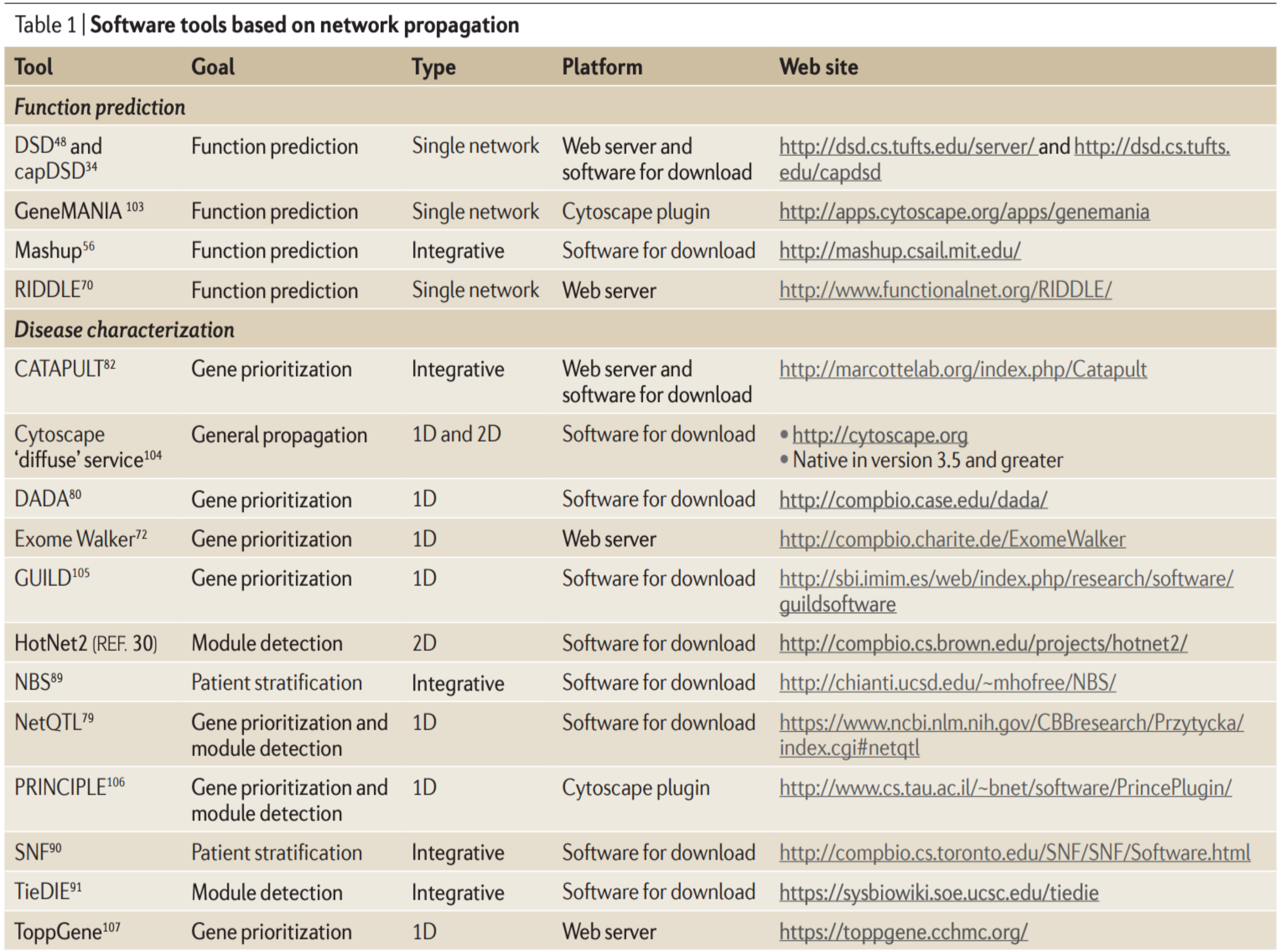


Table 1 | **Software tools based on network propagation**

Tool	Goal	Type	Platform	Web site
Function prediction				
DSD[48] and capDSD[34]	Function prediction	Single network	Web server and software for download	http://dsd.cs.tufts.edu/server/ and http://dsd.cs.tufts.edu/capdsd
GeneMANIA [103]	Function prediction	Single network	Cytoscape plugin	http://apps.cytoscape.org/apps/genemania
Mashup[56]	Function prediction	Integrative	Software for download	http://mashup.csail.mit.edu/
RIDDLE[70]	Function prediction	Single network	Web server	http://www.functionalnet.org/RIDDLE/
Disease characterization				
CATAPULT[82]	Gene prioritization	Integrative	Web server and software for download	http://marcottelab.org/index.php/Catapult
Cytoscape 'diffuse' service[104]	General propagation	1D and 2D	Software for download	• http://cytoscape.org • Native in version 3.5 and greater
DADA[80]	Gene prioritization	1D	Software for download	http://compbio.case.edu/dada/
Exome Walker[72]	Gene prioritization	1D	Web server	http://compbio.charite.de/ExomeWalker
GUILD[105]	Gene prioritization	1D	Software for download	http://sbi.imim.es/web/index.php/research/software/guildsoftware
HotNet2 (REF. 30)	Module detection	2D	Software for download	http://compbio.cs.brown.edu/projects/hotnet2/
NBS[89]	Patient stratification	Integrative	Software for download	http://chianti.ucsd.edu/~mhofree/NBS/
NetQTL[79]	Gene prioritization and module detection	1D	Software for download	https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#netqtl
PRINCIPLE[106]	Gene prioritization and module detection	1D	Cytoscape plugin	http://www.cs.tau.ac.il/~bnet/software/PrincePlugin/
SNF[90]	Patient stratification	Integrative	Software for download	http://compbio.cs.toronto.edu/SNF/SNF/Software.html
TieDIE[91]	Module detection	Integrative	Software for download	https://sysbiowiki.soe.ucsc.edu/tiedie
ToppGene[107]	Gene prioritization	1D	Web server	https://toppgene.cchmc.org/

Cowen *et al*, Nature Reviews Genetics (2017)

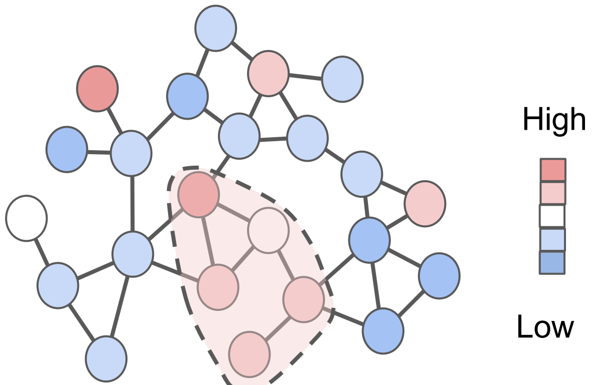
Existing algorithms do not have rigorous, theoretical guarantees

Most algorithms assess their performance using real biological datasets, e.g.

- Runtime
- Recovering known biological findings
- Discovery of potentially new biological insights

But **do not** assess performance on generative model of the data

-> obscures fundamental issues shared across algorithms



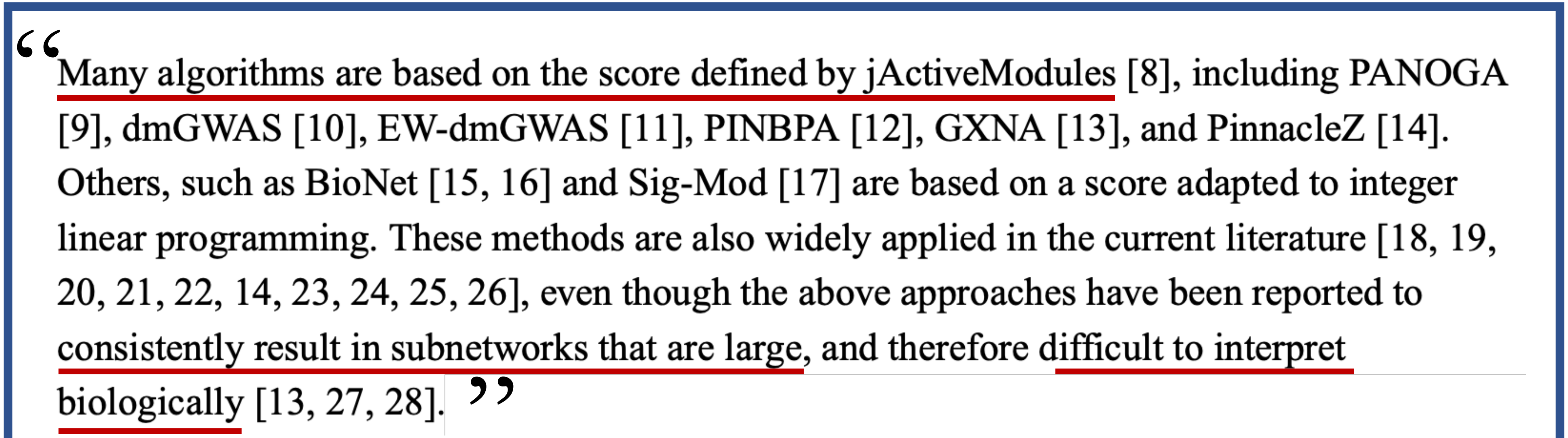
Altered Subnetwork Problem:

Given:

1) Network G = (V,E)
2) Vertex scores X_v (usually derived from p-values)

Goal: Identify high-scoring subnetworks H of G

Many algorithms output **very large subnetworks**



"Many algorithms are based on the score defined by jActiveModules [8], including PANOGA [9], dmGWAS [10], EW-dmGWAS [11], PINBPA [12], GXNA [13], and PinnacleZ [14]. Others, such as BioNet [15, 16] and Sig-Mod [17] are based on a score adapted to integer linear programming. These methods are also widely applied in the current literature [18, 19, 20, 21, 22, 14, 23, 24, 25, 26], even though the above approaches have been reported to consistently result in subnetworks that are large, and therefore difficult to interpret biologically [13, 27, 28]."

"Network module identification—a widespread theoretical bias and best practices" by Nikolayeva *et al* (Methods 2018)

Altered Subnetwork Problem:

Given:

1) Network G = (V,E)
2) Vertex scores X_v (usually derived from p-values)

Goal: Identify high-scoring subnetworks H of G

jActiveModules/Cytoscape (Ideker et al, 2002): maximizes function over connected subgraphs

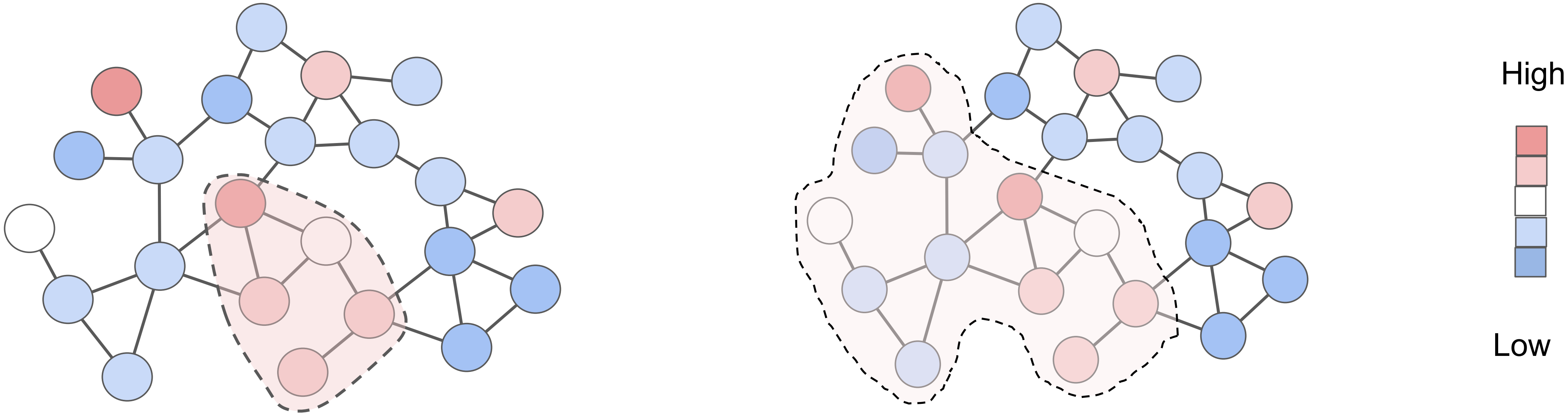
$$\arg\max_{\text{connected } S} \frac{1}{\sqrt{|S|}} \sum_{v\in S} X_v$$

A simple simulation with an implanted subnetwork

Network has **10,000** vertices, implanted altered subnetwork A has **500** vertices

- Vertex scores in A are ~2 standard deviations larger than avg

jActiveModules outputs a subnetwork with **2505** vertices (5x increase!)



True

Estimated

Many **heuristics** for reducing size – but effectiveness is unclear

Our contributions

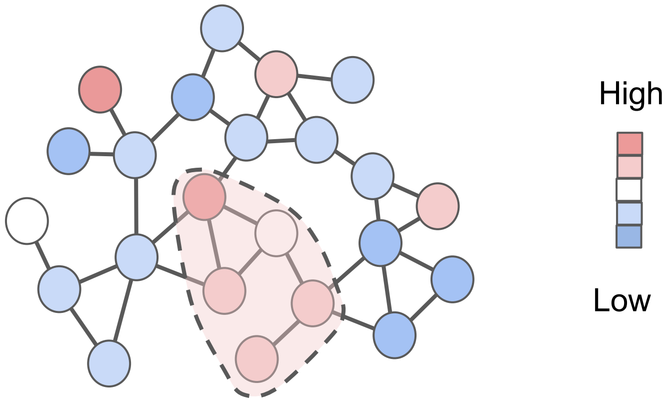
1. **Generative model** for altered subnetworks

2. Show issue of identifying large subnetworks is due to **statistical bias**

3. Develop NetMix algorithm which is **asymptotically unbiased**

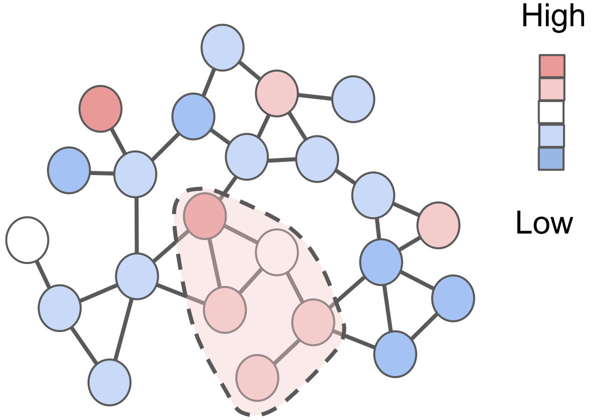
Extensions:

- NetMix2 algorithm which uses network propagation (random walks)
- Anomaly detection in statistics/ML



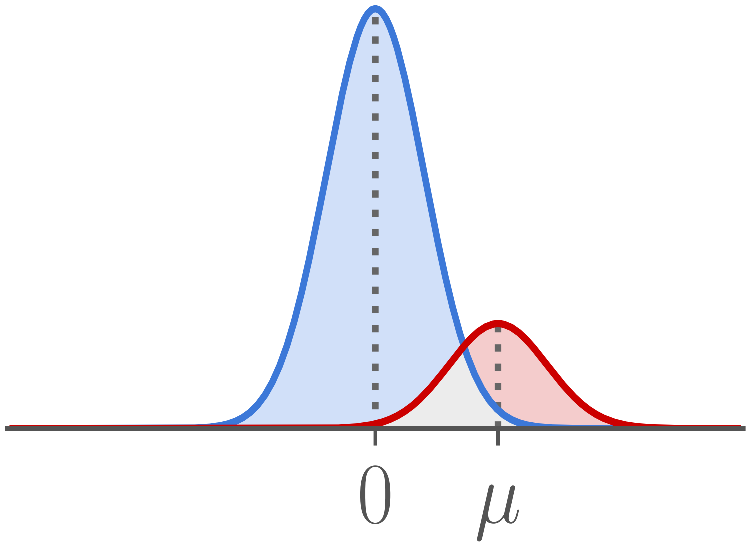
Generative model: Altered Subnetwork Distribution

- G=(V, E) is a graph
- A ⊆ V is a connected subgraph, or the altered subnetwork



Vertex scores $(X_v)_{v \in V}$ are distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$



Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v \in V}$ distributed as

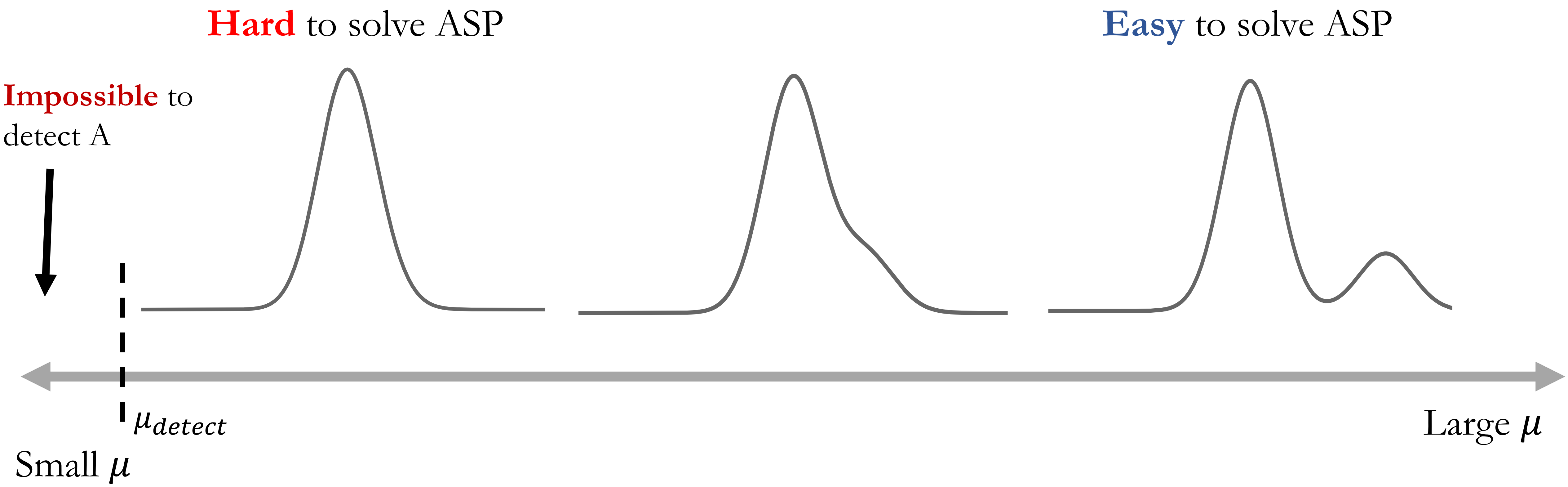
$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0,1) & \text{otherwise} \end{cases}$$

find the altered subnetwork A.

Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v\in V}$ distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

find the altered subnetwork A.



Hard to solve ASP

Easy to solve ASP

Impossible to detect A

μ_{detect}

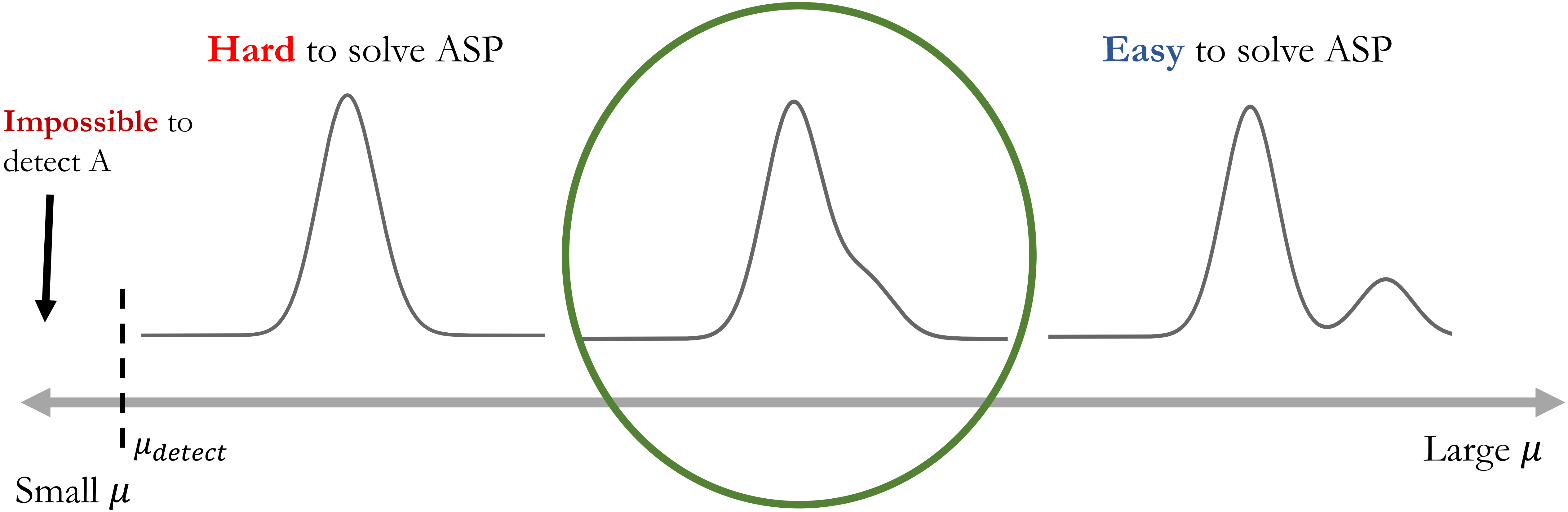
Small μ

Large μ

Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v\in V}$ distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

find the altered subnetwork A.



Hard to solve ASP

Easy to solve ASP

Impossible to detect A

μ_{detect}

Small μ

Large μ

Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v \in V}$ distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0,1) & \text{otherwise} \end{cases}$$

find the altered subnetwork A.

Theorem: **Maximum Likelihood Estimator (MLE)** of the altered subnetwork A is:

$$\widehat{A}_{\text{MLE}} = \underset{\substack{S \subseteq V \\ S \text{ connected}}}{\operatorname{argmax}} \left(\frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v \right)$$

Altered Subnetwork Problem: Given graph G and vertex scores $(X_v)_{v \in V}$ distributed as

$$X_v \sim \begin{cases} N(\mu, 1) & \text{if } v \in A \\ N(0,1) & \text{otherwise} \end{cases}$$

find the altered subnetwork A.

Theorem: **Maximum Likelihood Estimator (MLE)** of the altered subnetwork A is:

$$\widehat{A}_{\text{MLE}} = \underset{\substack{S \subseteq V \\ S \text{ connected}}}{\operatorname{argmax}} \left(\frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v \right)$$

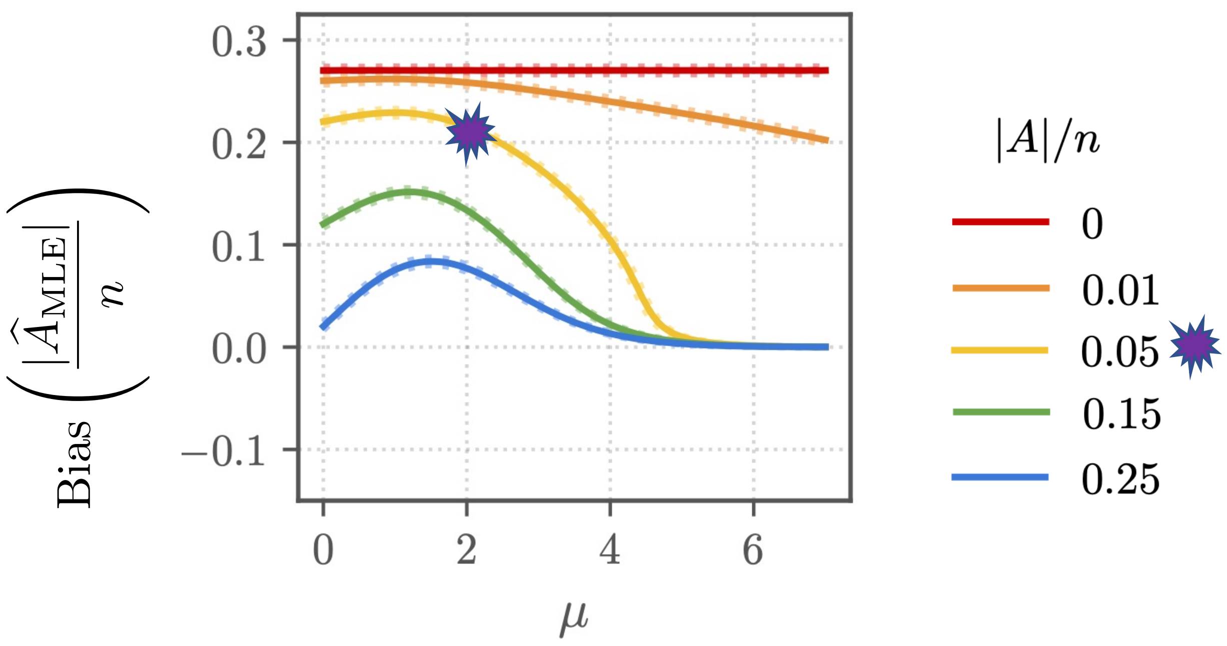
MLE = jActiveModules!

jActiveModules paper (Ideker et al, 2002) does not describe generative model nor the connection to the MLE

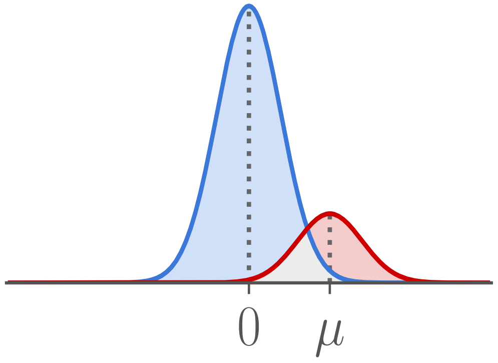
MLE is biased estimator

$$\boxed{\text{Bias}\left(\frac{|\widehat{A}_{\text{MLE}}|}{n}\right) \triangleq E\left[\frac{|\widehat{A}_{\text{MLE}}|}{n}\right] - \frac{|A|}{n}}$$

We observe that MLE has **positive bias:** MLE overestimates the size $\boldsymbol{\frac{|A|}{n}}$ of the altered subnetwork on average (where n=|V|)



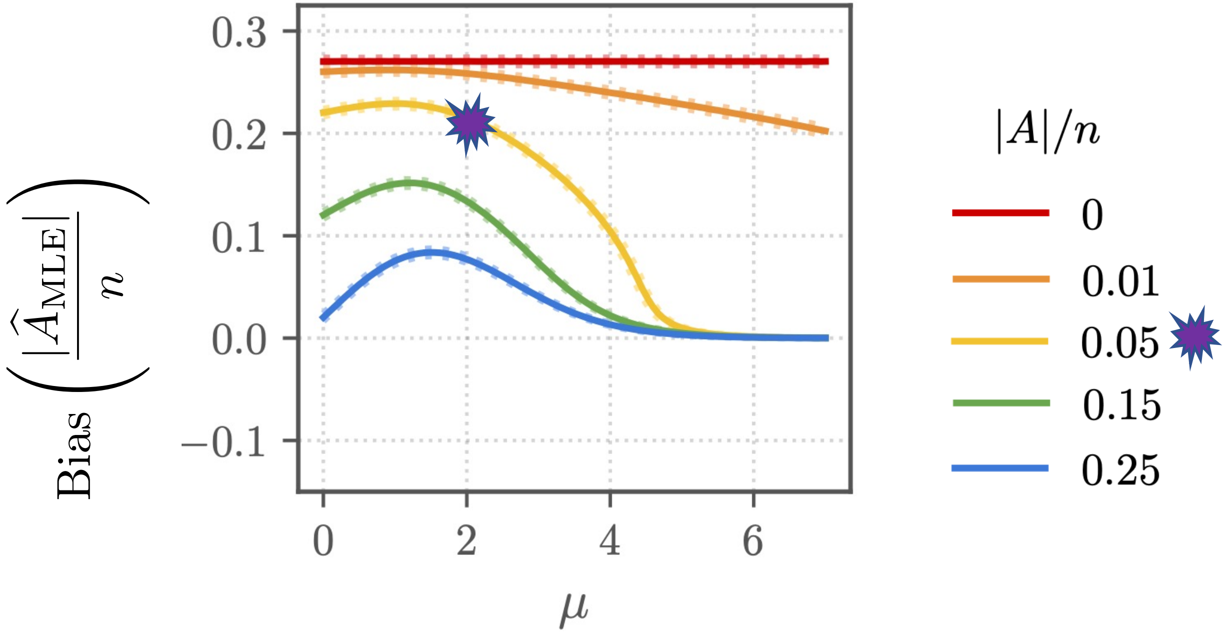
(Simulation from before)



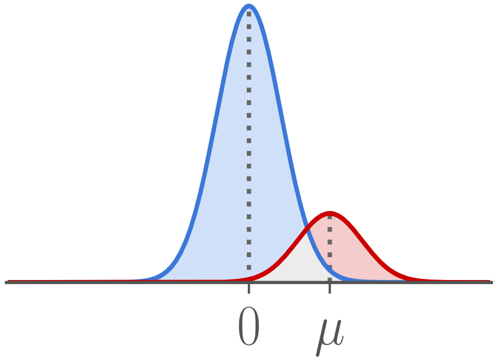
MLE is biased estimator

$$\boxed{\text{Bias}\left(\frac{|\widehat{A}_{\text{MLE}}|}{n}\right) \triangleq E\left[\frac{|\widehat{A}_{\text{MLE}}|}{n}\right] - \frac{|A|}{n}}$$

We observe that MLE has **positive bias:** MLE overestimates the size $\frac{|A|}{n}$ of the altered subnetwork on average (where n=|V|)



(Simulation from before)



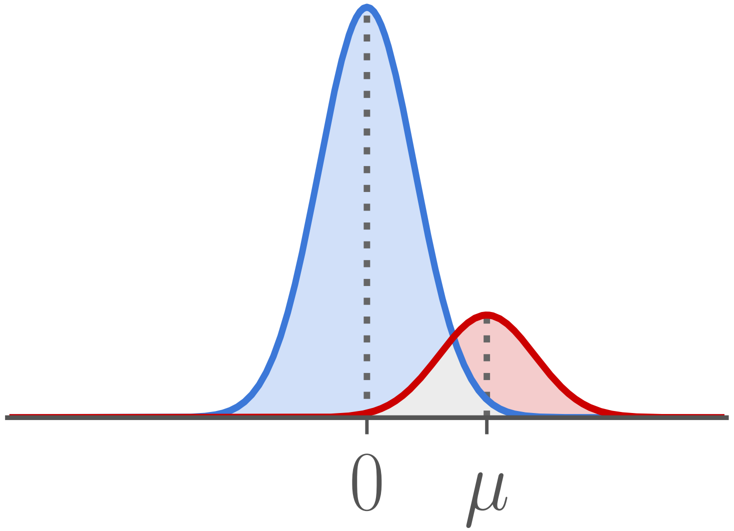
We prove MLE is <u>asymptotically biased:</u>

$$\lim_{n\to\infty} \text{Bias}\left(\frac{|\widehat{A}_{\text{MLE}}|}{n}\right) > 0$$

Assuming number of connected subgraphs is exponential
(RECOMB 2020; ICML 2021; unpublished w/ H Schmidt)

How to reduce bias?

Key idea: Model the distribution of the vertex scores before using the network



Fit vertex scores to **Gaussian Mixture Model (GMM):**

$$X_v \sim (1-\alpha)\cdot N(0,1) + \alpha \cdot N(\mu, 1)$$

α = proportion of vertices in altered subnetwork
μ = mean of altered subnetwork distribution

GMM yields less biased estimate of altered subnetwork size

MLE: $\widehat{A}_{\mathrm{MLE}} = \underset{\substack{S \subseteq V \\ S \text{ connected}}}{\operatorname{argmax}} \left(\frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v \right)$

vs

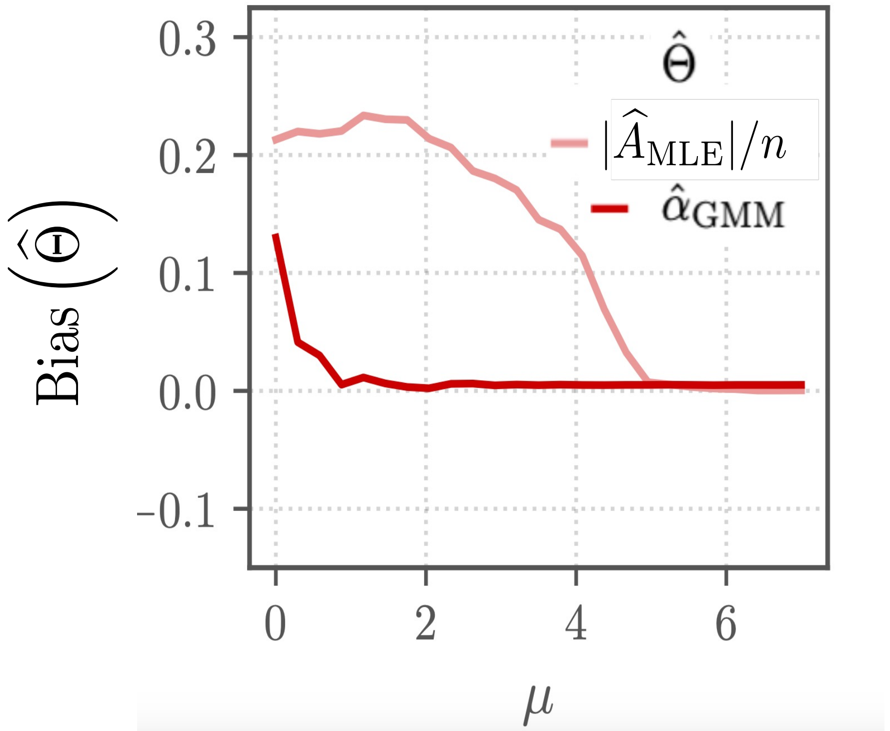
GMM: Fit vertex scores X_v to GMM

$$X_v \sim (1-\alpha) \cdot N(0,1) + \alpha \cdot N(\mu, 1)$$

and estimate GMM parameters $\widehat{\alpha}_{\mathrm{GMM}}, \widehat{\mu}_{\mathrm{GMM}}$

α = proportion of vertices in altered subnetwork

μ = mean of altered subnetwork distribution



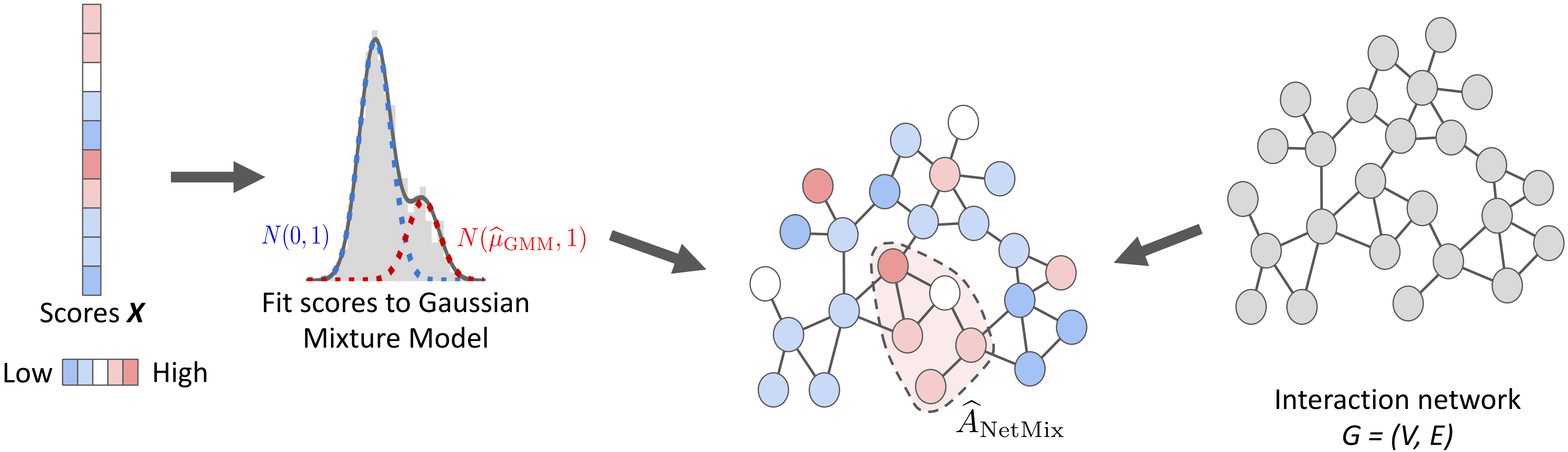
We prove: GMM estimator is asymptotically unbiased (ICML 2021)

-> **Model mis-specification** helps!
(Fitting ASD with GMM)

NetMix Algorithm

Given vertex scores $(X_v)_{v \in V}$ and graph G:

1. Fit scores to GMM using EM, and compute *responsibilities* $r_v = P(v \in A \mid X_v)$
2. Find connected subnetwork $\widehat{A}_{\mathrm{NetMix}}$ with GMM-estimated size and largest total responsibility


$N(0,1)$
$N(\widehat{\mu}_{\mathrm{GMM}},1)$
$\widehat{A}_{\mathrm{NetMix}}$

Scores ***X***

Low 🟦🟦⬜🟥🟥 High

Fit scores to Gaussian Mixture Model

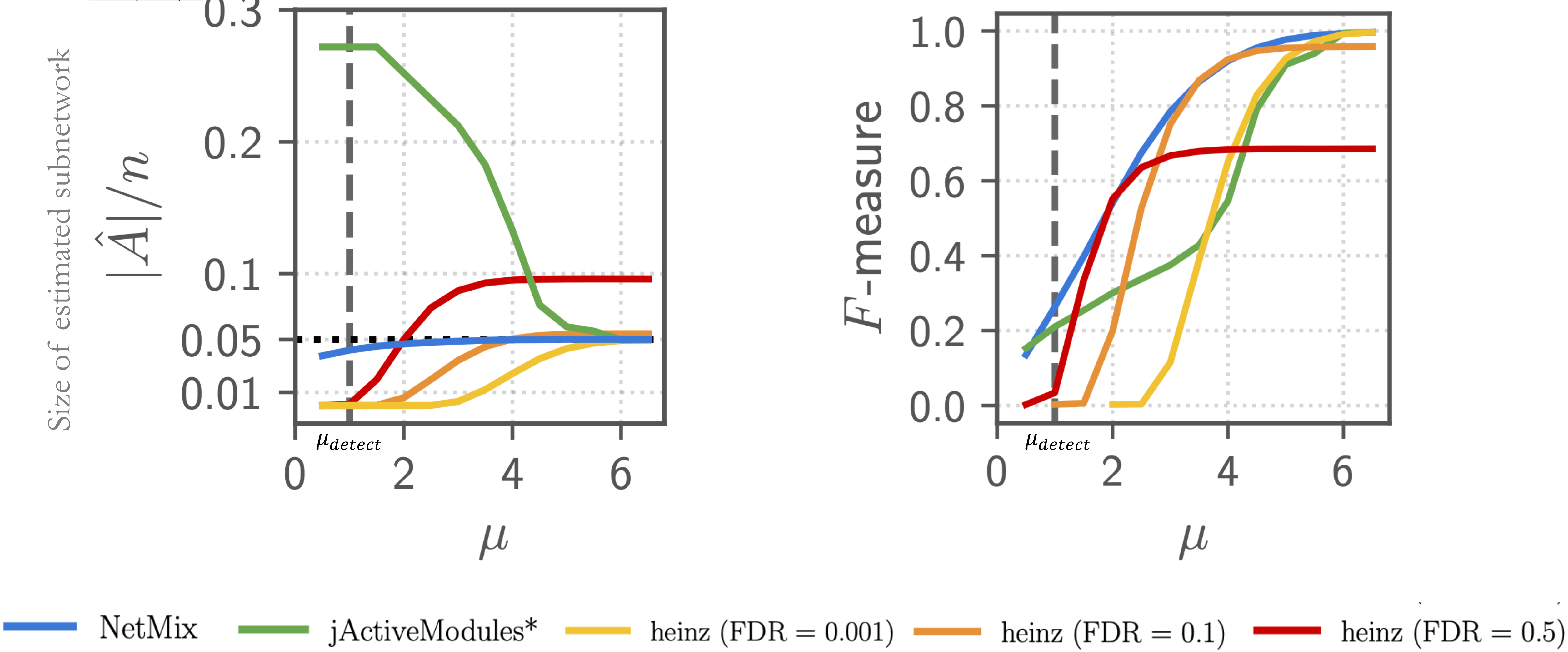
Identify connected subnetwork with largest total responsibility (ILP)

Interaction network

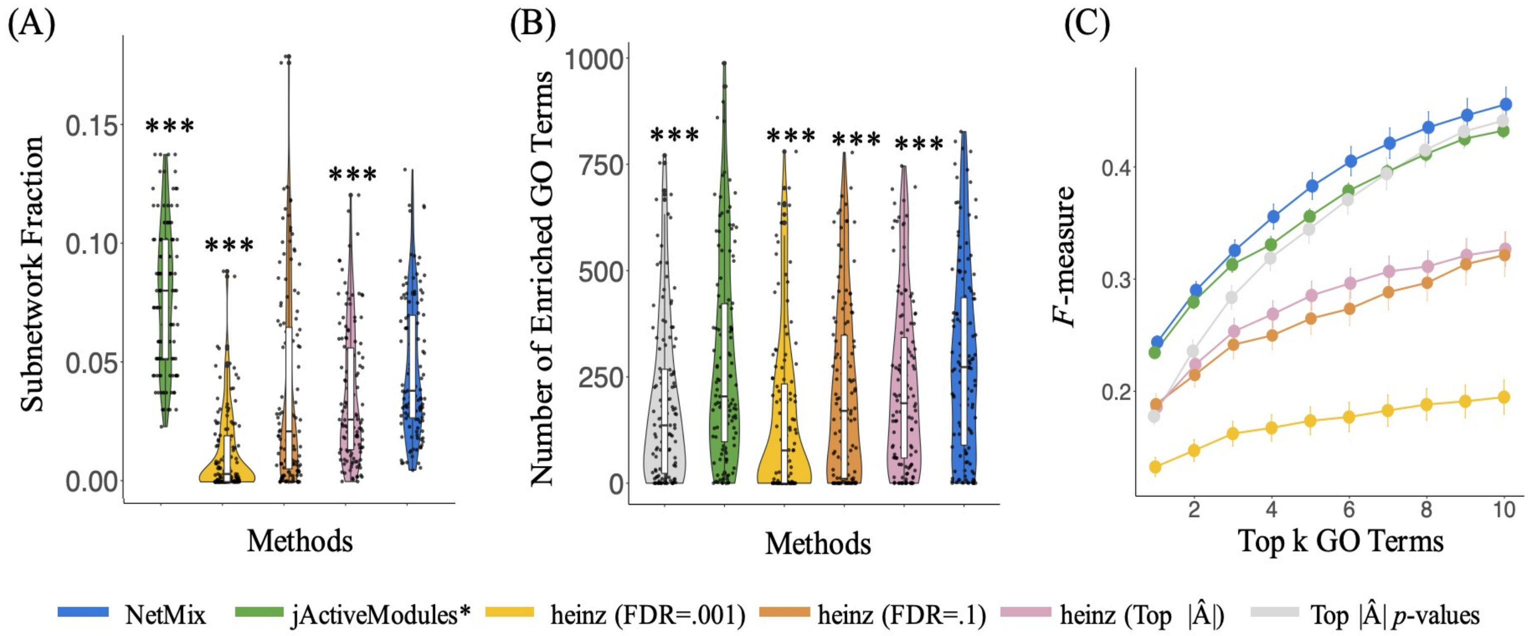
G = (V, E)

Results – simulated data

Real PPI network ($n \approx$ 15,000 vertices)
Implanted altered subnetwork A
w/ size $|A| = 0.05n = 750$



Results – differential gene expression + somatic mutations in cancer



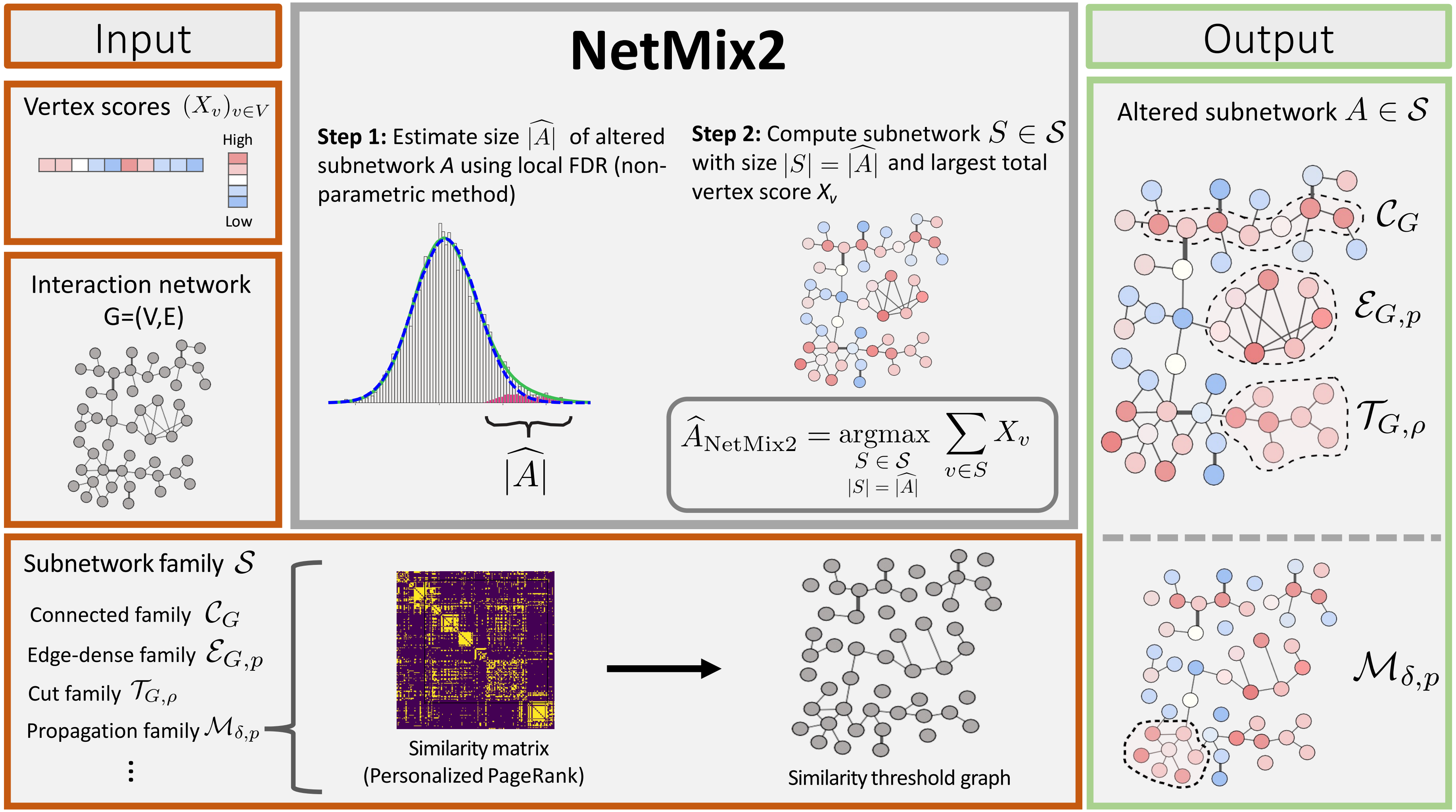
157 gene expression experiments from Expression Atlas (Petryszak et al, 2015)

	Network			
Method	None	HINT+HI	iRefIndex	ReactomeFI
jActiveModules*	2,136 / 0.155	1,575 / 0.191	1,815 / 0.174	557 / 0.261
jActiveModules (Greedy search)	N.A. / N.A.	N.A. / N.A.	N.A. / N.A.	N.A. / N.A.
jActiveModules (Simulated annealing)	N.A. / N.A.	12,284 / 0.086	15,046 / 0.074	8,329 / 0.118
heinz (FDR = 0.001)	115 / 0.205	119 / 0.216	109 / 0.217	114 / 0.215
heinz (FDR = 0.1)	259 / 0.244	249 / 0.264	259 / 0.255	253 / 0.215
Hierarchical Hotnet	N.A. / N.A.	228 / 0.214	297 / 0.215	228 / 0.214
NetMix	307 / **0.254**	263 / **0.277**	296 / **0.270**	264 / **0.270**

Cancer driver gene prediction:

- Using MutSigCV2 *p*-values and multiple interaction networks

NetMix2: extension to other distributions and graph topologies



Input

Vertex scores $(X_v)_{v\in V}$

High

Low

Interaction network G=(V,E)

Subnetwork family $\mathcal{S}$

Connected family $\mathcal{C}_G$

Edge-dense family $\mathcal{E}_{G,p}$

Cut family $\mathcal{T}_{G,\rho}$

Propagation family $\mathcal{M}_{\delta,p}$

⋮

Similarity matrix (Personalized PageRank)

Similarity threshold graph

NetMix2

Step 1: Estimate size $|\widehat{A}|$ of altered subnetwork *A* using local FDR (non-parametric method)

$|\widehat{A}|$

Step 2: Compute subnetwork $S \in \mathcal{S}$ with size $|S| = |\widehat{A}|$ and largest total vertex score X_v

$$\widehat{A}_{\mathrm{NetMix2}} = \underset{\substack{S\in\mathcal{S}\\ |S|=|\widehat{A}|}}{\operatorname{argmax}} \sum_{v\in S} X_v$$

Output

Altered subnetwork $A \in \mathcal{S}$

$\mathcal{C}_G$

$\mathcal{E}_{G,p}$

$\mathcal{T}_{G,\rho}$

$\mathcal{M}_{\delta,p}$

Contributions:

1. **Non-parametric** estimation of altered subnetwork size

2. Different **subnetwork topologies** (connectivity, edge density, cut size, …)
   - Define topology for **network propagation** (random walks)

In paper (RECOMB 2022 + JCB): improved identification of disease genes in **cancer + GWAS**

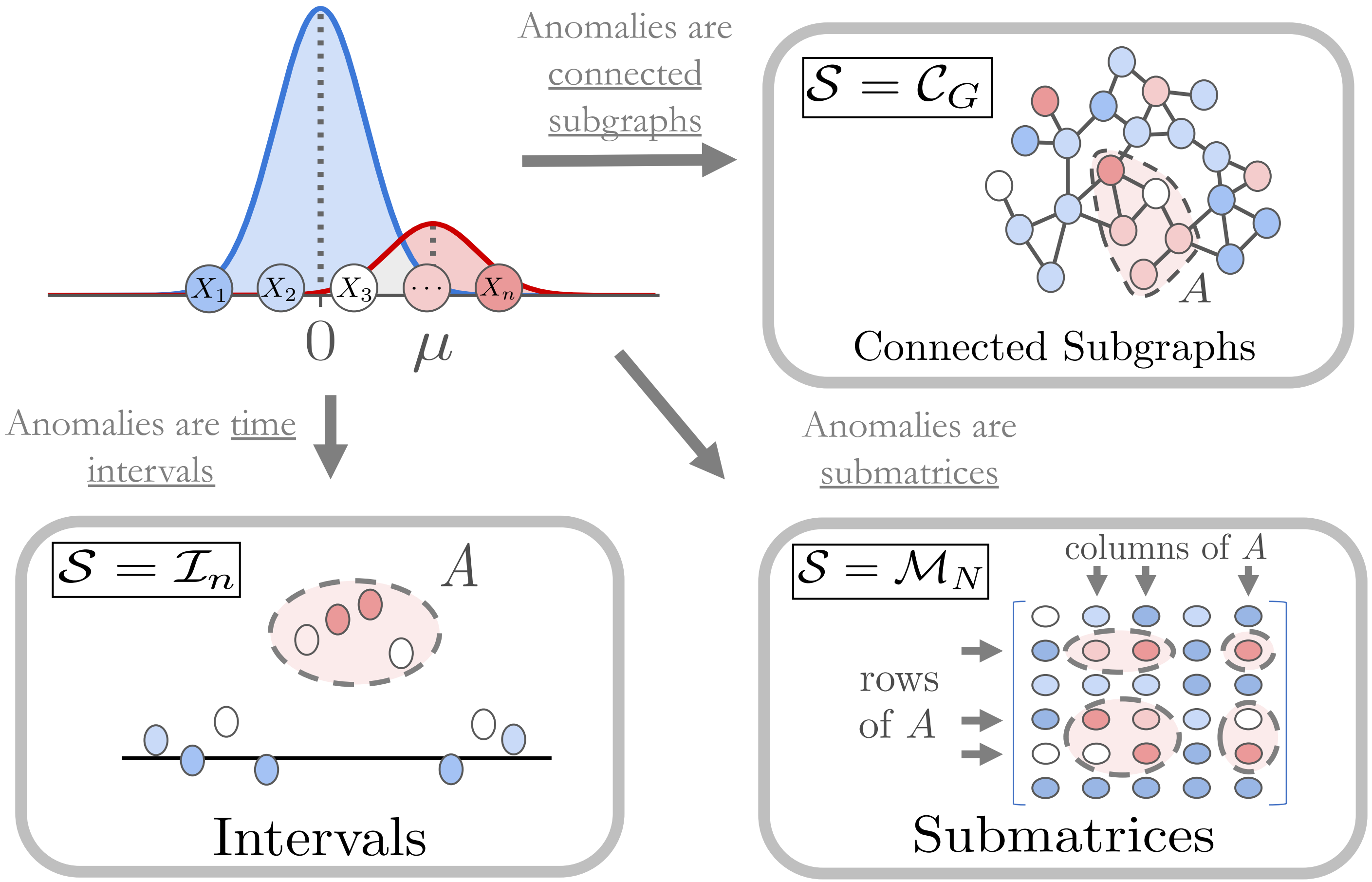
Anomaly detection in statistics/ML

Normal means problem: Data $X_1, \ldots, X_n$ independently distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0,1) & \text{otherwise} \end{cases}$$

where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$

ICML 2021: **We extend theoretical results** and show: MLE is biased iff number of sets in anomaly family $\mathcal{S}$ containing A is exponential



Anomalies are connected subgraphs

$\mathcal{S} = \mathcal{C}_G$

Connected Subgraphs

Anomalies are time intervals

$\mathcal{S} = \mathcal{I}_n$

Intervals

Anomalies are submatrices

$\mathcal{S} = \mathcal{M}_N$

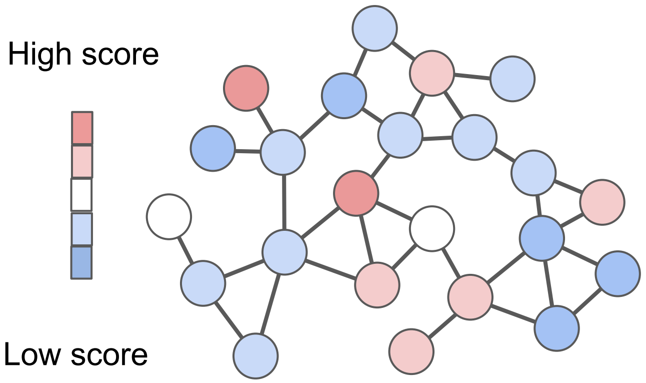
columns of A

rows of A

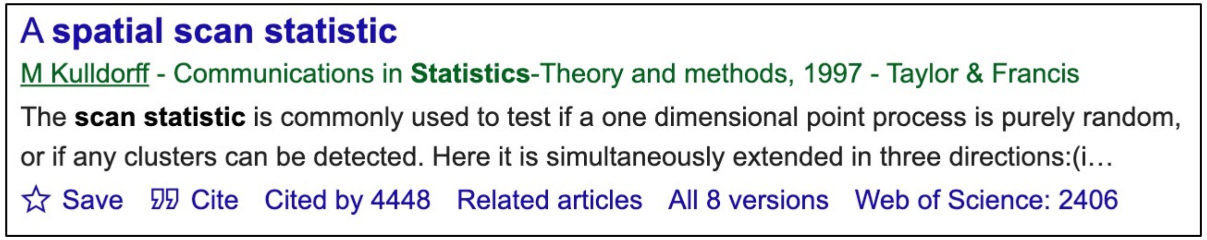
Submatrices

MLE is biased for spatial anomalies

Spatial adjacency graph



- Vertices = points in space
- Edges connect adjacent points in space
- Score = disease incidence



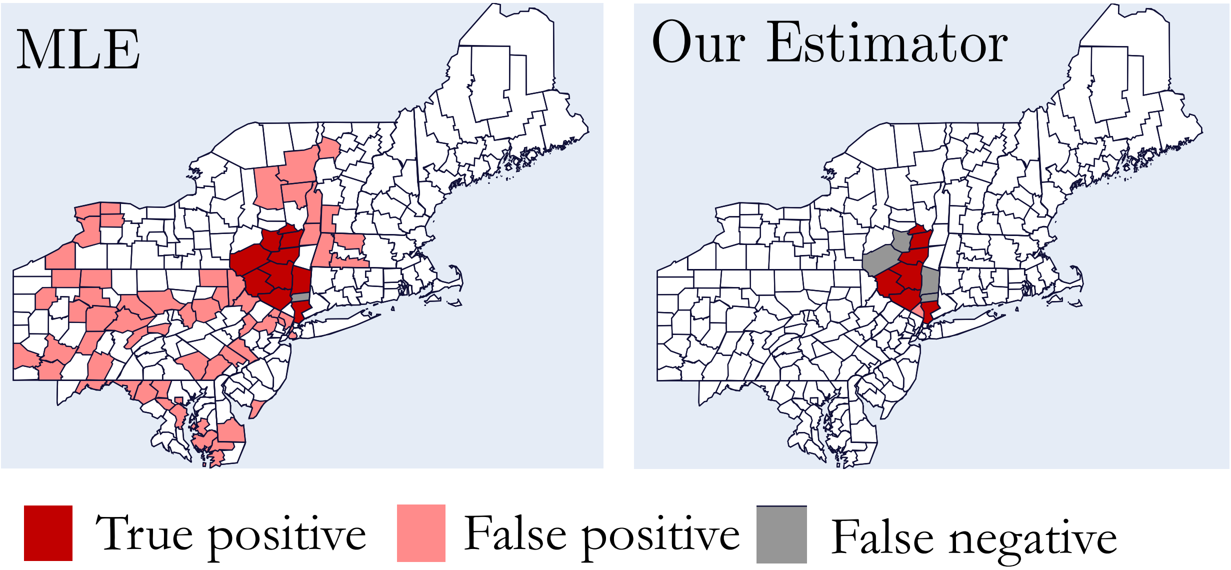
A spatial scan statistic

M Kulldorff - Communications in **Statistics**-Theory and methods, 1997 - Taylor & Francis

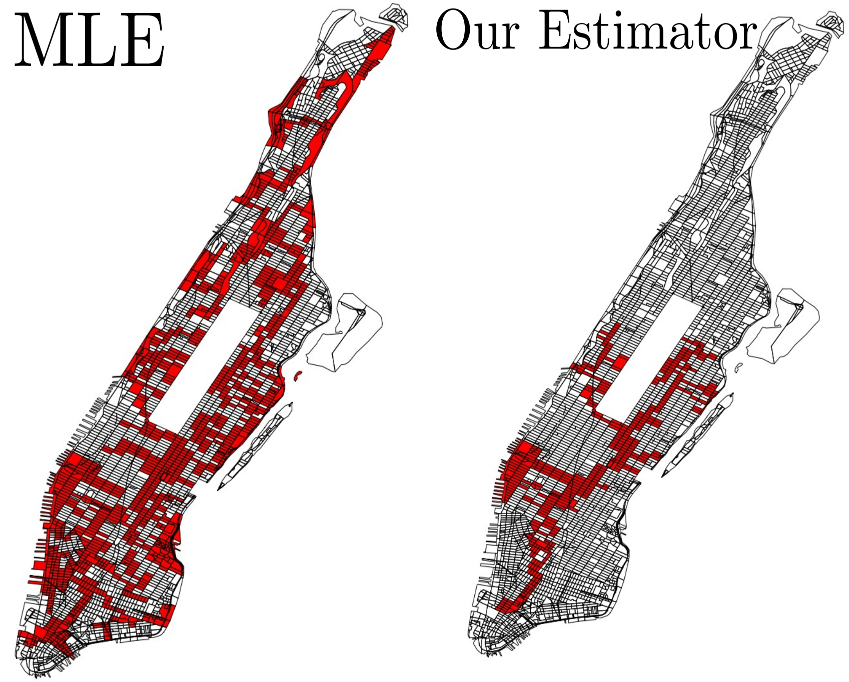
The **scan statistic** is commonly used to test if a one dimensional point process is purely random, or if any clusters can be detected. Here it is simultaneously extended in three directions:(i…

☆ Save 99 Cite Cited by 4448 Related articles All 8 versions Web of Science: 2406

MLE = network version of widely-used *"spatial scan statistic"*



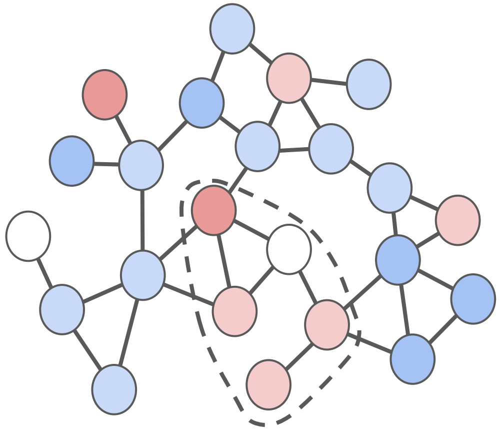
Simulated disease outbreak in northeast US



Real data: NYC breast cancer incidence

Summary

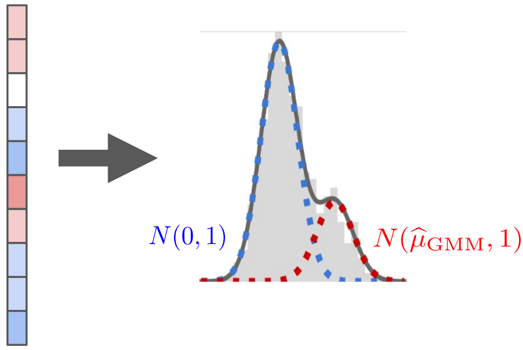
Generative model for altered subnetworks



<u>**We show:**</u> MLE is asymptotically biased for connected subgraphs

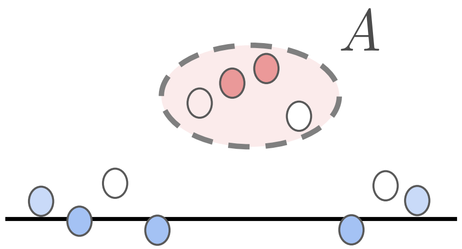
NetMix/NetMix2: asymptotically unbiased altered subnetwork algorithms

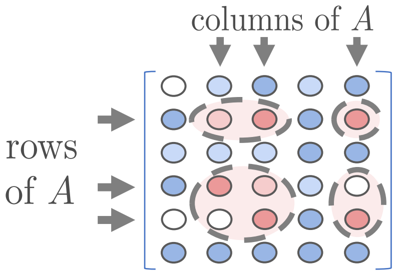
<u>Idea:</u> fit vertex scores to mixture model **before** using network



Results extend to **anomaly detection** in machine learning/statistics

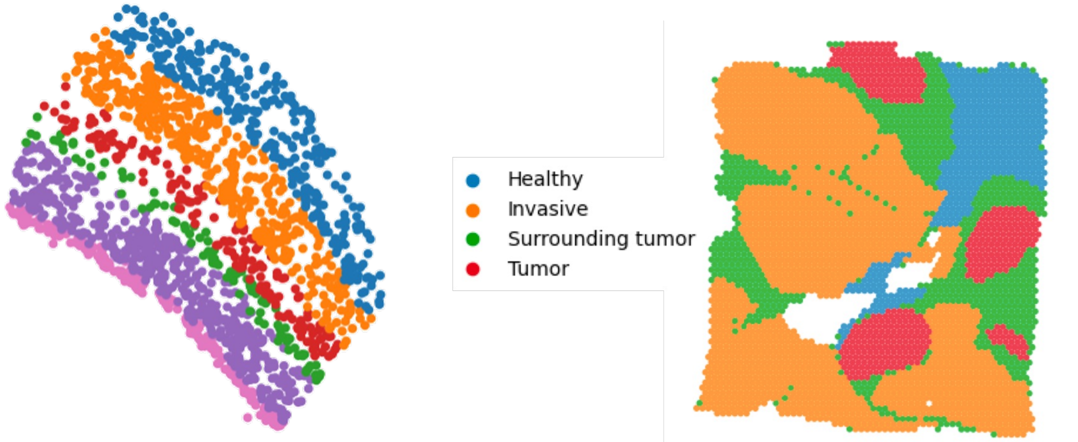
Future idea: anomaly detection in spatial transcriptomics?





My thesis: computational methods for understanding complex biological systems

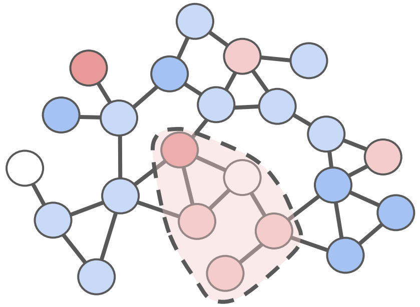
Spatial biology

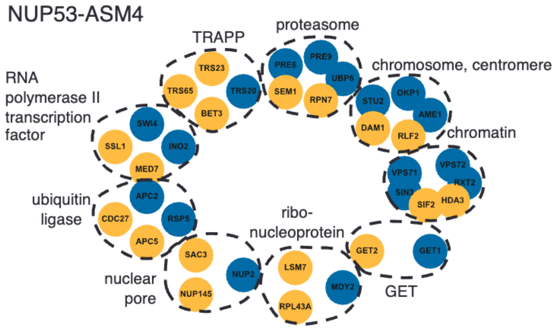


Spatial variation in gene expression

- Ma*, **Chitra***, et al. *RECOMB 2022 + Cell Systems.*
- **Chitra** et al. *RECOMB 2024 + in review at Nature Methods.*

Network interactions and anomalies





Altered subnetwork identification

- Reyna*, **Chitra***, et al. *RECOMB 2020 + JCB.*
- **Chitra** et al. *ICML 2021.*
- **Chitra***, Park*, Raphael. *RECOMB 2022 + JCB.*

Cell-cell interactions

- Sarkar*, **Chitra***, et al. *In submission at ISMB 2024.*

Learning genetic interactions

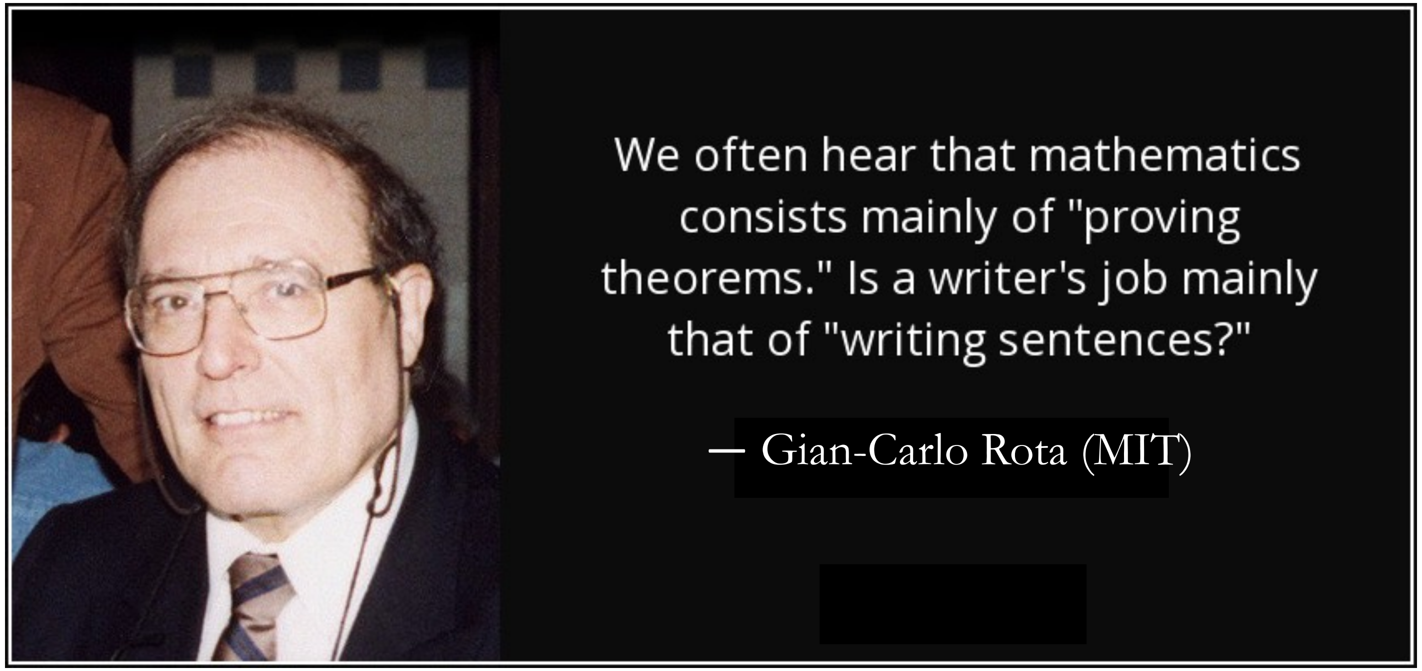
- **Chitra***, Arnold*, Raphael. *In review at Nature Genetics.*
- Shuaibi*, **Chitra***, Raphael. *In submission at RECOMB-CCB.*

Machine learning + data mining

- **Chitra** and Raphael. *ICML 2019.*
- **Chitra** and Musco. *WSDM 2020.*

* indicates joint first authorship

What does a computational biology researcher do?



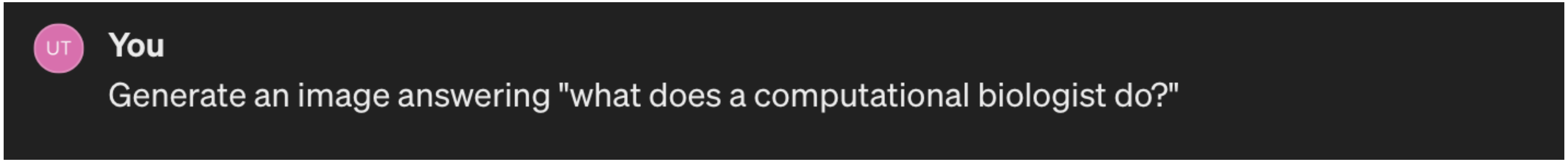
We often hear that mathematics consists mainly of "proving theorems." Is a writer's job mainly that of "writing sentences?"

— Gian-Carlo Rota (MIT)

[redacted]

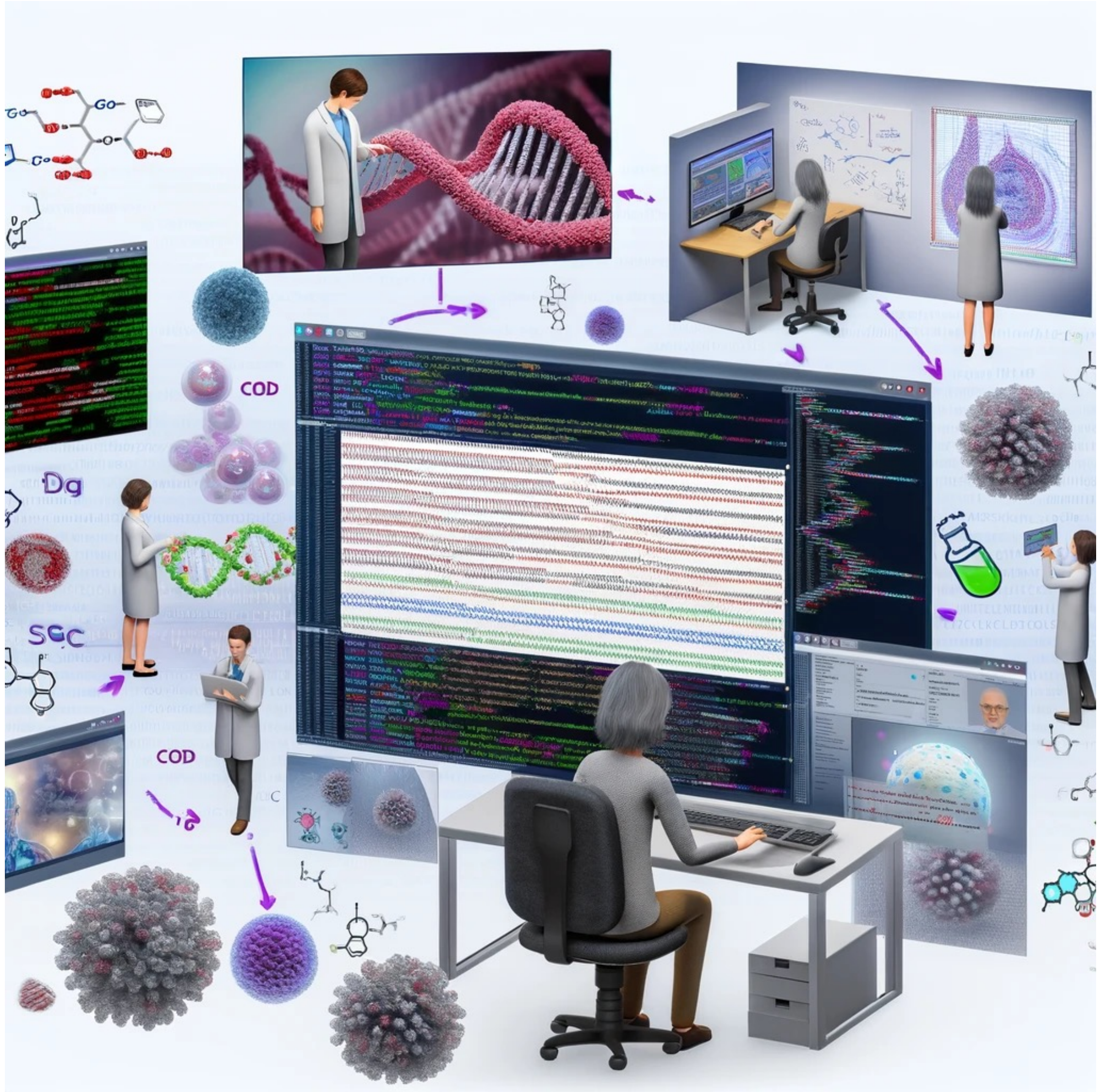
Is a computational biologist's job mainly that of "analyzing biological data?"

No! We also try to (1) identify biologically interesting problems and (2) find mathematically "elegant" solutions to problems



You

Generate an image answering "what does a computational biologist do?"



Thank you!

Ben Raphael

Advisor

Committee

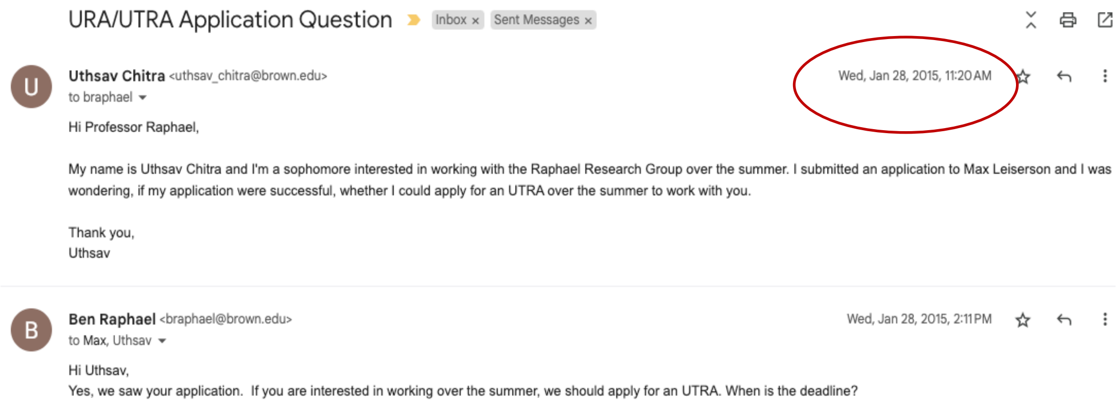
Bernard Chazelle

Ellen Zhong

Yuri Pritykin

Fei Chen

Almost a decade working with Ben



URA/UTRA Application Question

Uthsav Chitra <uthsav_chitra@brown.edu>
to braphael

Wed, Jan 28, 2015, 11:20 AM

Hi Professor Raphael,

My name is Uthsav Chitra and I'm a sophomore interested in working with the Raphael Research Group over the summer. I submitted an application to Max Leiserson and I was wondering, if my application were successful, whether I could apply for an UTRA over the summer to work with you.

Thank you,
Uthsav

Ben Raphael <braphael@brown.edu>
to Max, Uthsav

Wed, Jan 28, 2015, 2:11 PM

Hi Uthsav,
Yes, we saw your application. If you are interested in working over the summer, we should apply for an UTRA. When is the deadline?

Group retreat, summer 2015

Group retreat, summer 2023

Acknowledgments

Collaborators/co-authors:
Ben Raphael
Matt Reyna
Rebecca Elyanow
Kimberly Ding
Jasper Lee
Tyler Park
Cong Ma
Shirley Zhang
Brian Arnold
Sereno Lopez-Darwin
Hirak Sarkar
Kohei Sanno
Ahmed Shuaibi
Julian Gold
Clover Zheng
Sunay Joshi
Chris Musco
Tarun Chitra

Raphael group (past and present):
Simone Zaccaria
Ron Zeira
Pijus Simonaitis
Gryte Satas
Matt Myers
Hongyu Zheng
Palash Sashittal
Uyen Mai
Metin Balaban
Richard Zhang
Alexander Strzalkowski
Henri Schmidt
Xinhao Liu
Akhil Jakatdar
Gary Hu
Peter Halmos
Gillian Chu
Clover Zheng
Maya Gupta
Madelyne Xiao

+ support of numerous friends + family

NSF GRADUATE RESEARCH FELLOWSHIP PROGRAM
National Human Genome Research Institute
Siebel Scholars

Extra content

Belayer – simulated data

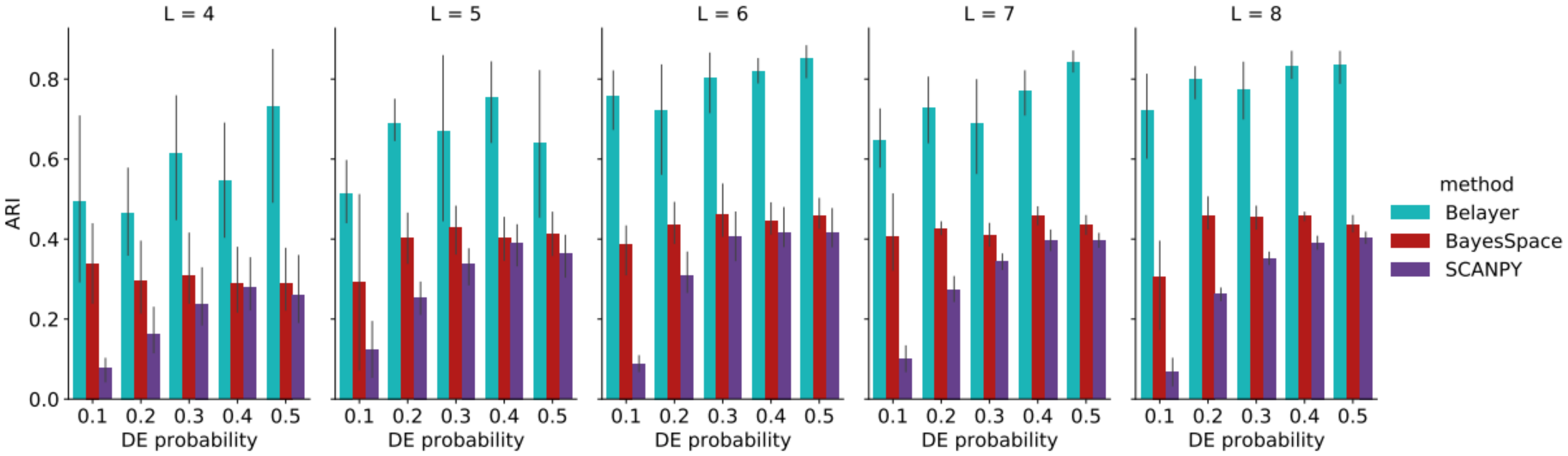
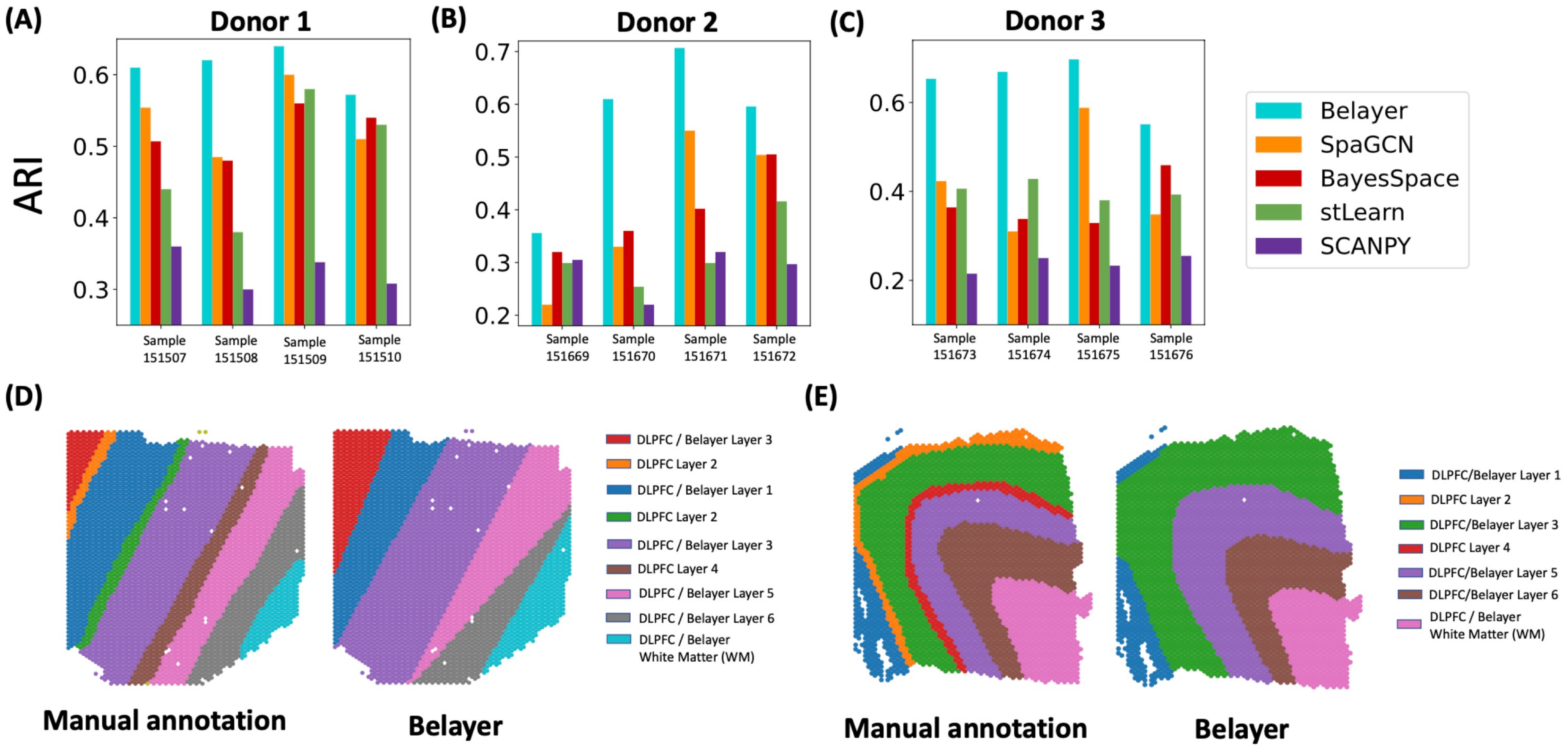


Figure S6: Comparison of Belayer, BayesSpace, and SCANPY in identifying spatially distinct cell clusters in the second simulation. Performance of each method is evaluated according to the Adjusted Rand Index (ARI) and shown for different values of the number L of layers and differential expression (DE) probability. Error bars indicate variation from 5 randomly simulated datasets for each parameter setting.

Belayer accurately identifies cortical layers in human DLPFC



Takeaway: ***Global*** *spatial model (Belayer) >* ***local*** *models*

Different accuracy metrics: DLPFC

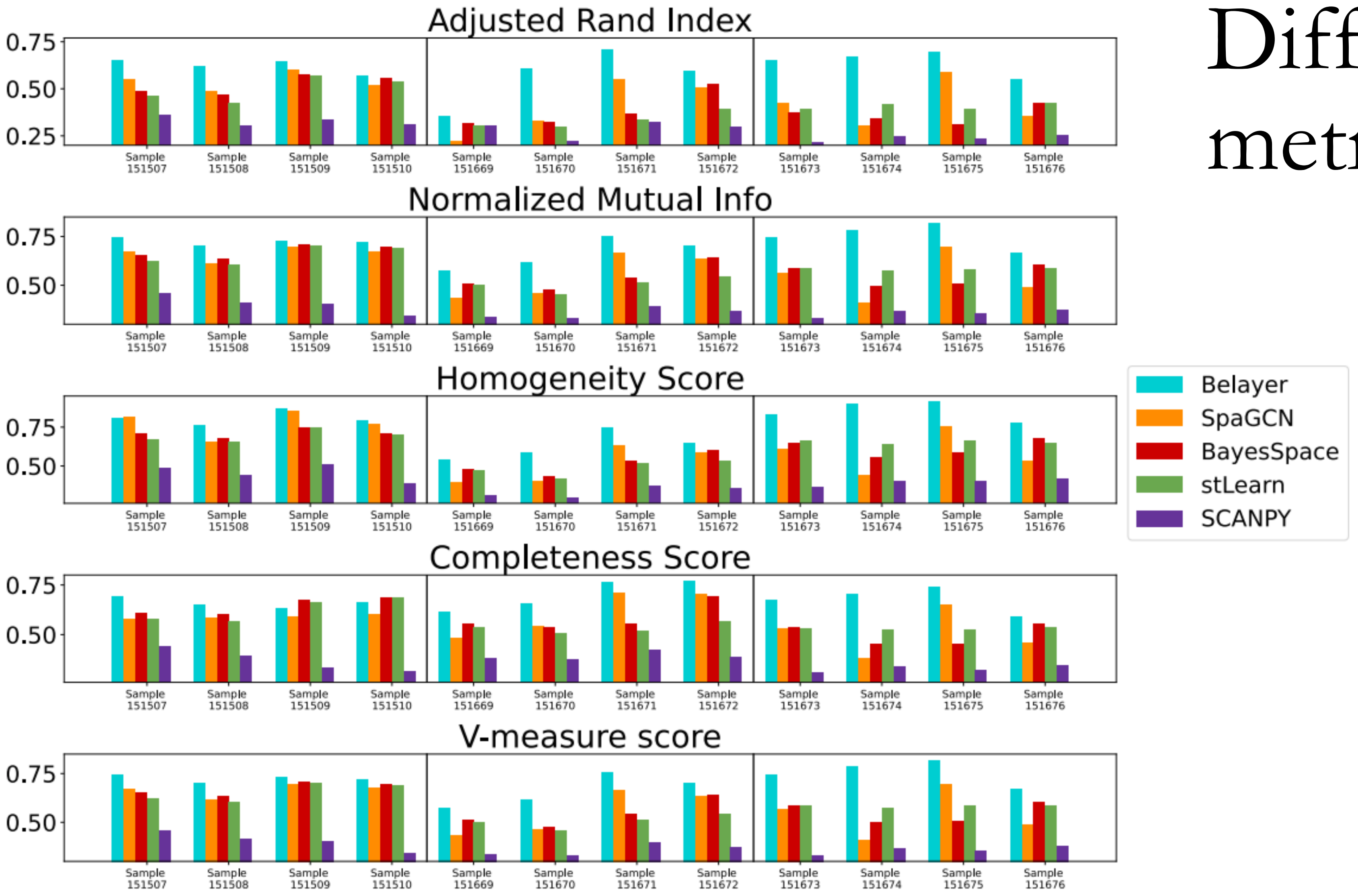
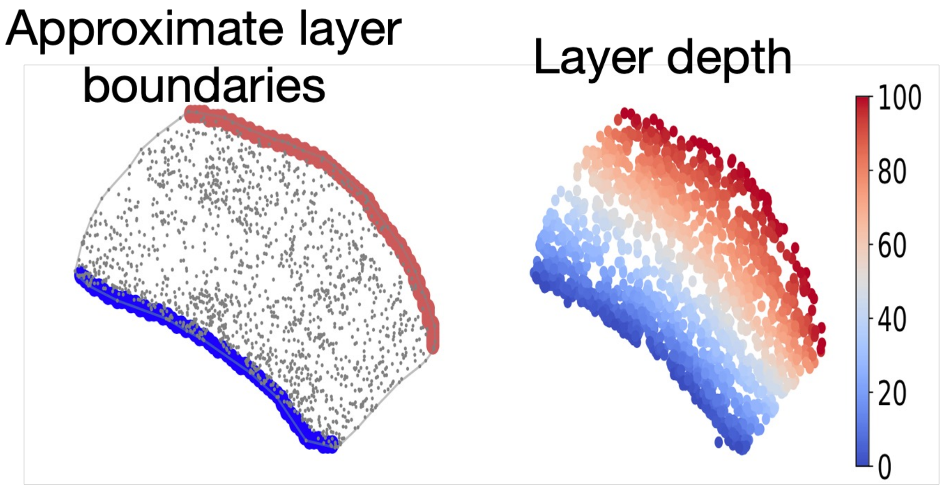


Figure S9: Comparison of Belayer and other methods from Figure 3 with different metrics.

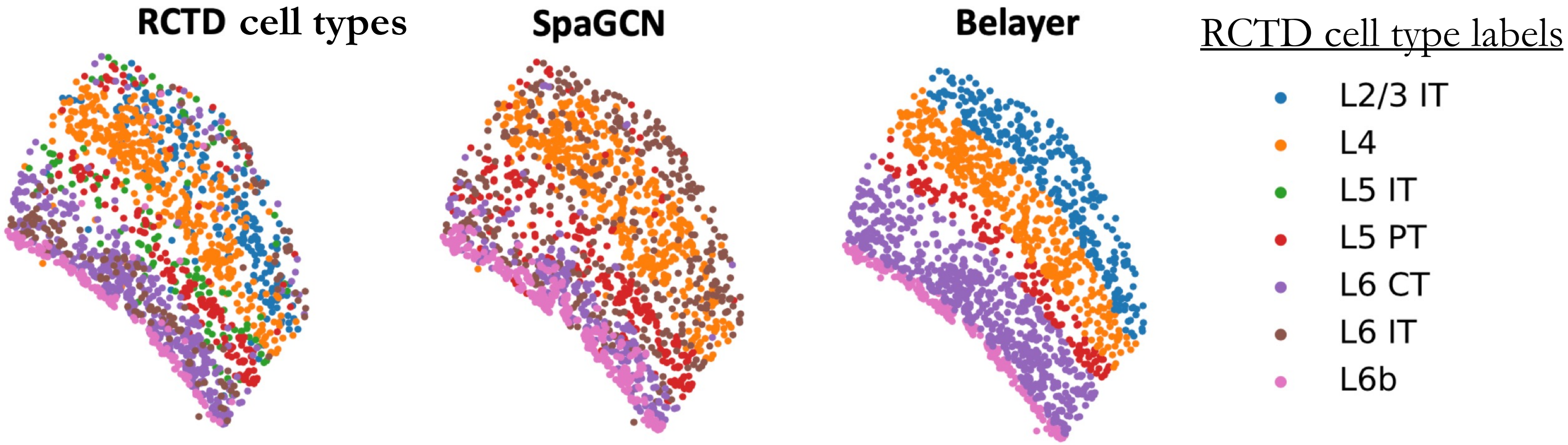
Belayer identifies spatially coherent cortical layers

(mouse somatosensory cortex data, SlideSeqV2)



SpaGCN: reference-free + **local** spatial model (GNN)

Belayer: reference-free + **global** spatial model



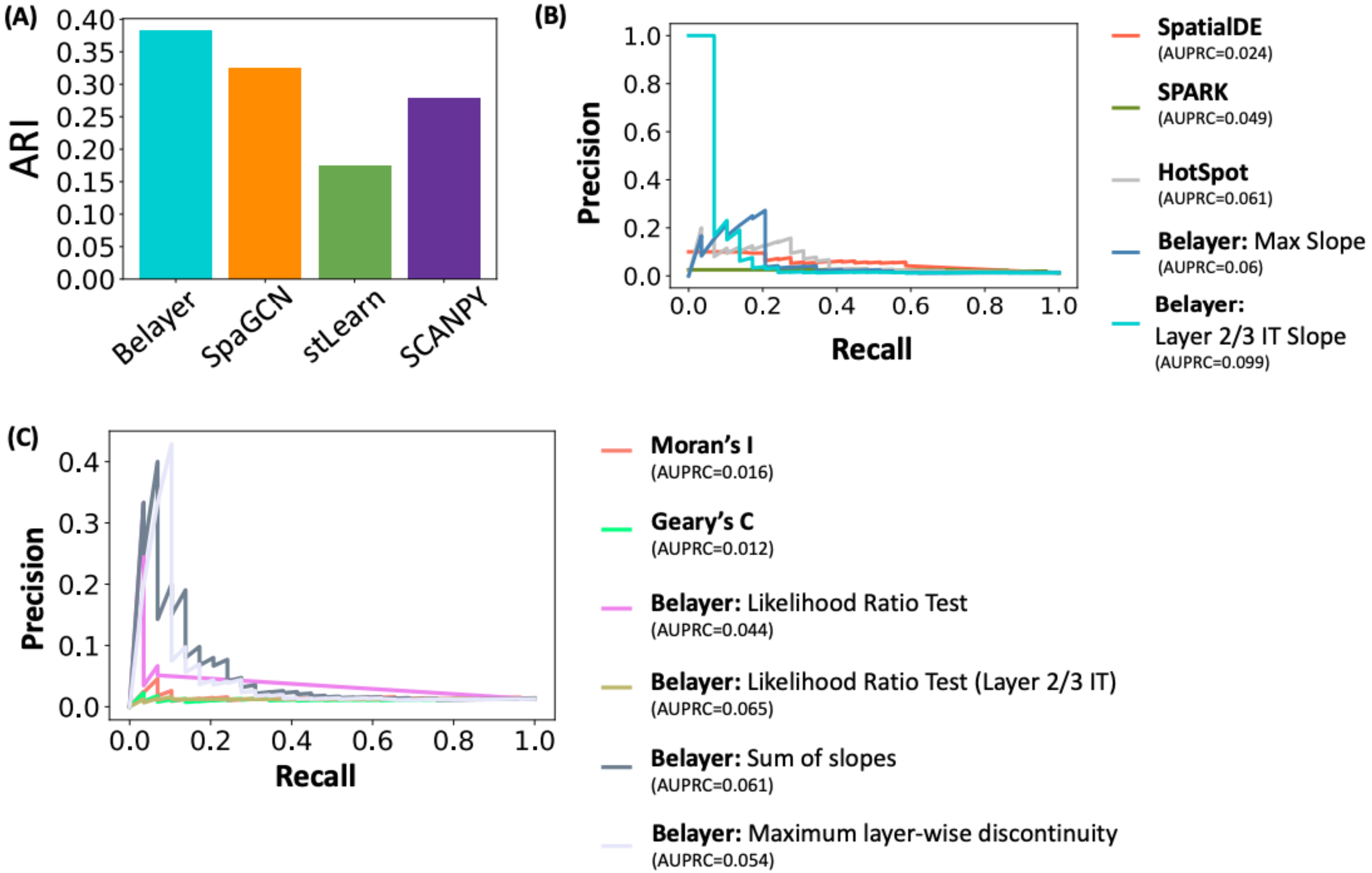
ARI=0.34

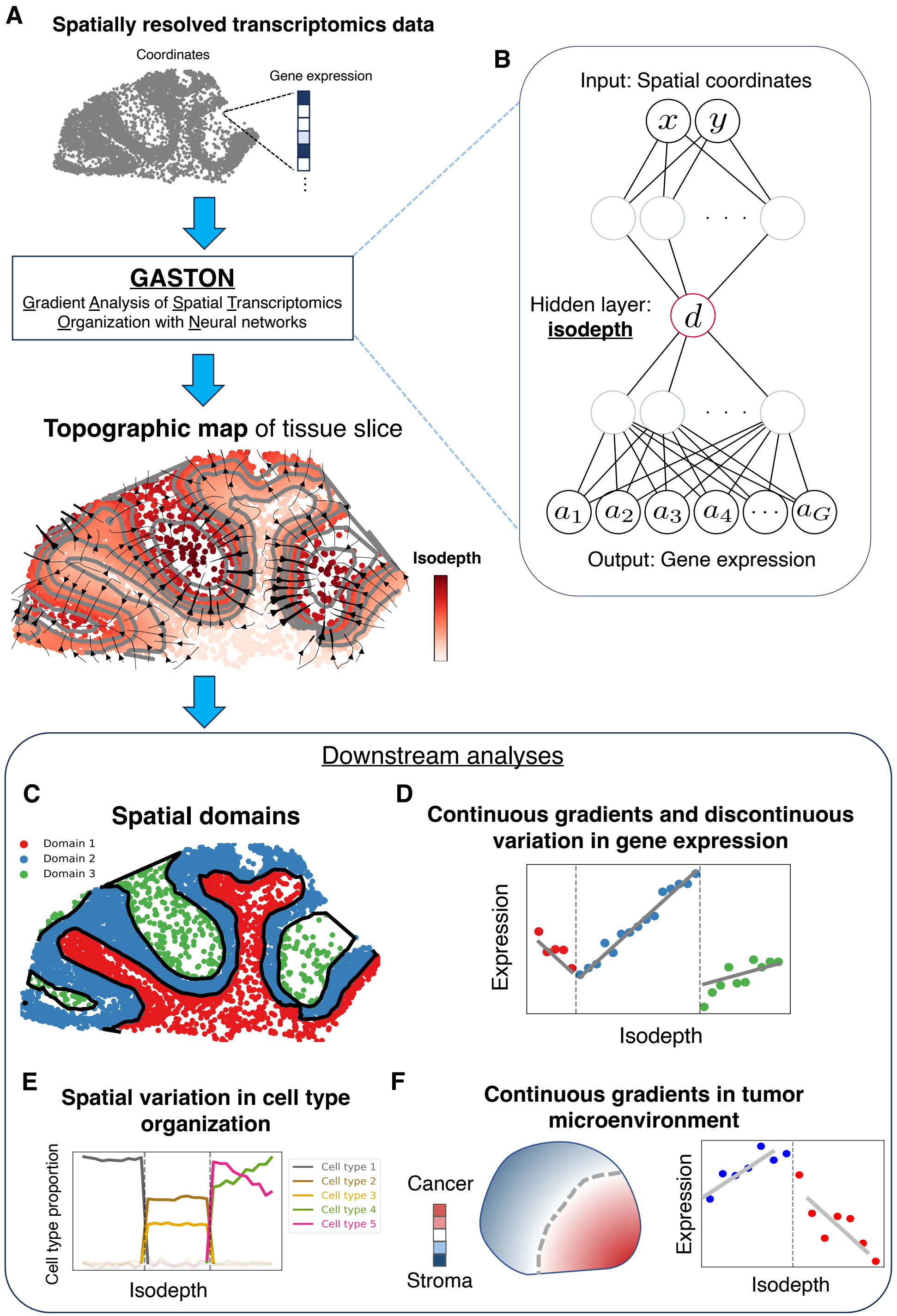
ARI=0.37

Data from Stickels et al., 2021
RCTD [Cable et al., 2021]

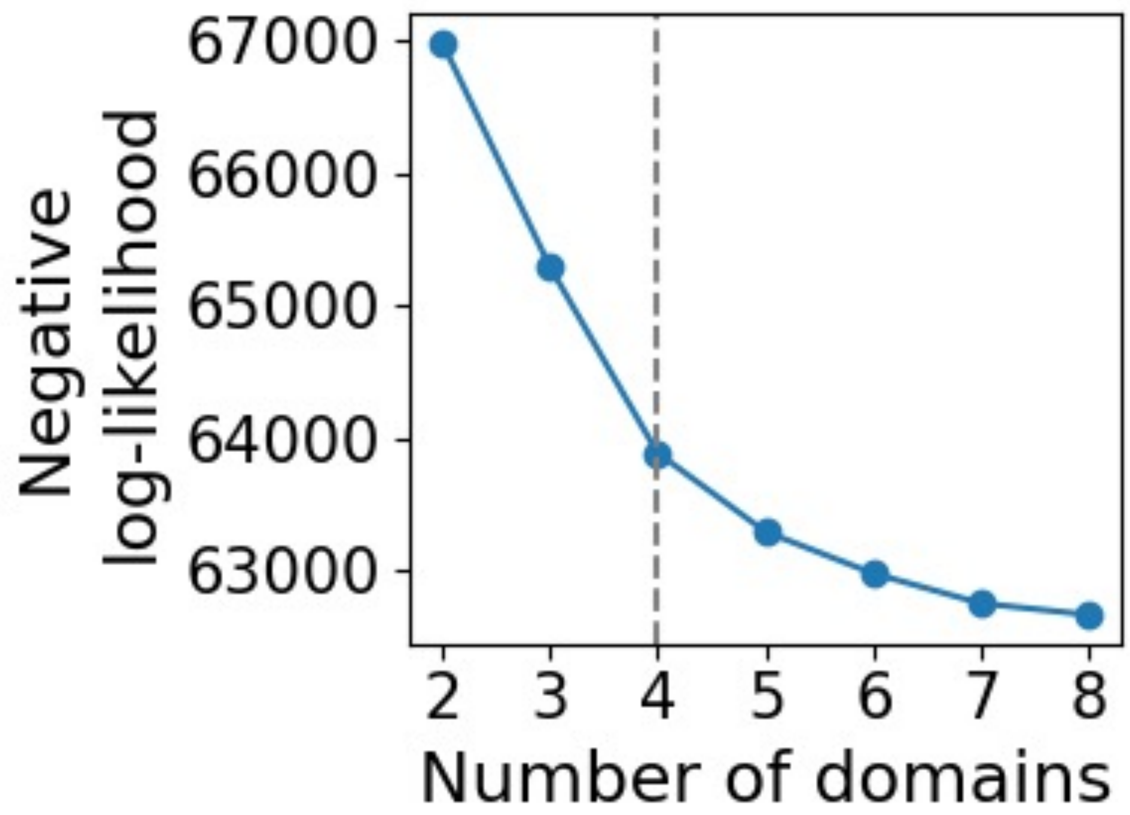
Belayer identifies spatially varying genes

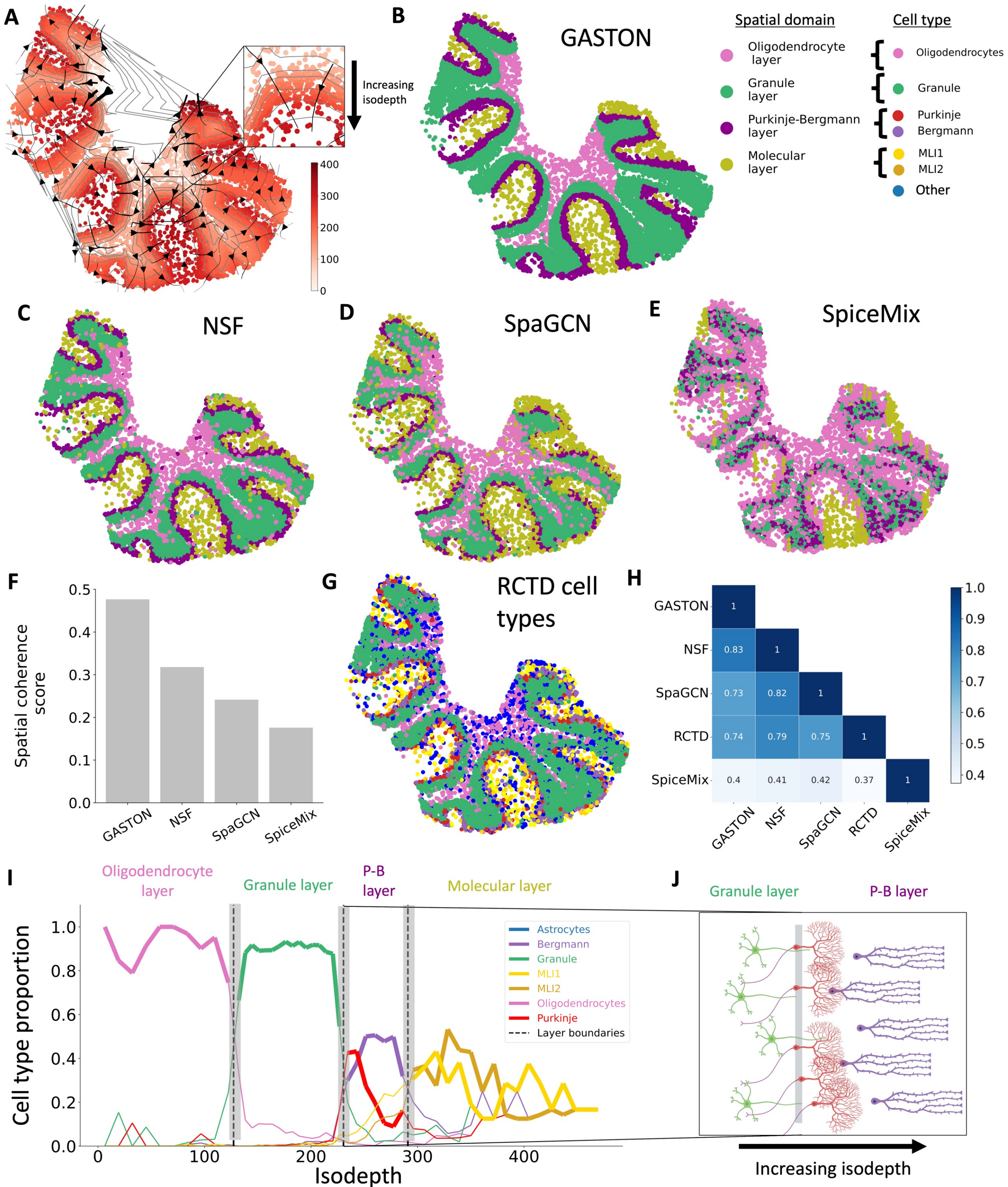
(mouse somatosensory cortex data, SlideSeqV2)



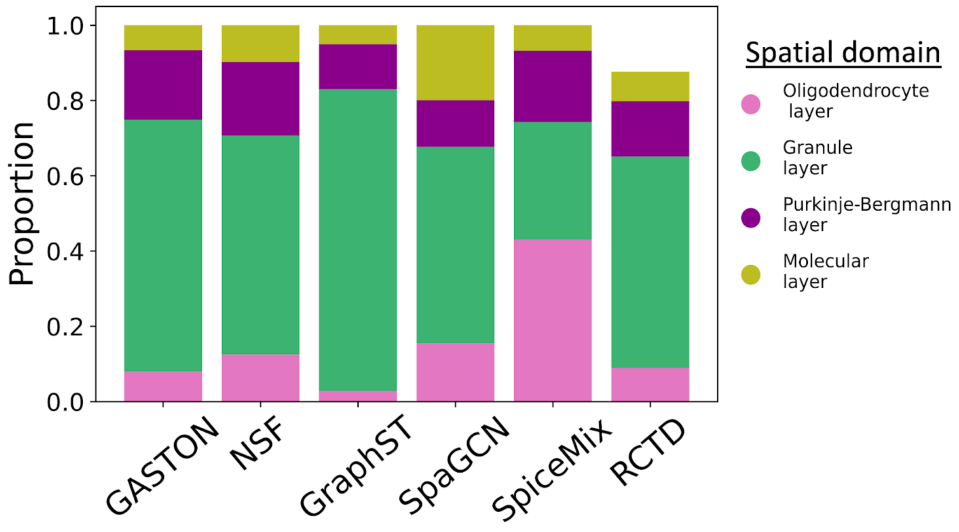

A
Spatially resolved transcriptomics data
Coordinates
Gene expression
GASTON
Gradient Analysis of Spatial Transcriptomics Organization with Neural networks
Topographic map of tissue slice
Isodepth
B
Input: Spatial coordinates
x
y
Hidden layer: isodepth
d
a_1
a_2
a_3
a_4
$\cdots$
a_G
Output: Gene expression
Downstream analyses
C
Spatial domains
Domain 1
Domain 2
Domain 3
D
Continuous gradients and discontinuous variation in gene expression
Expression
Isodepth
E
Spatial variation in cell type organization
Cell type proportion
Isodepth
Cell type 1
Cell type 2
Cell type 3
Cell type 4
Cell type 5
F
Continuous gradients in tumor microenvironment
Cancer
Stroma
Expression
Isodepth

GASTON – model selection (elbow)

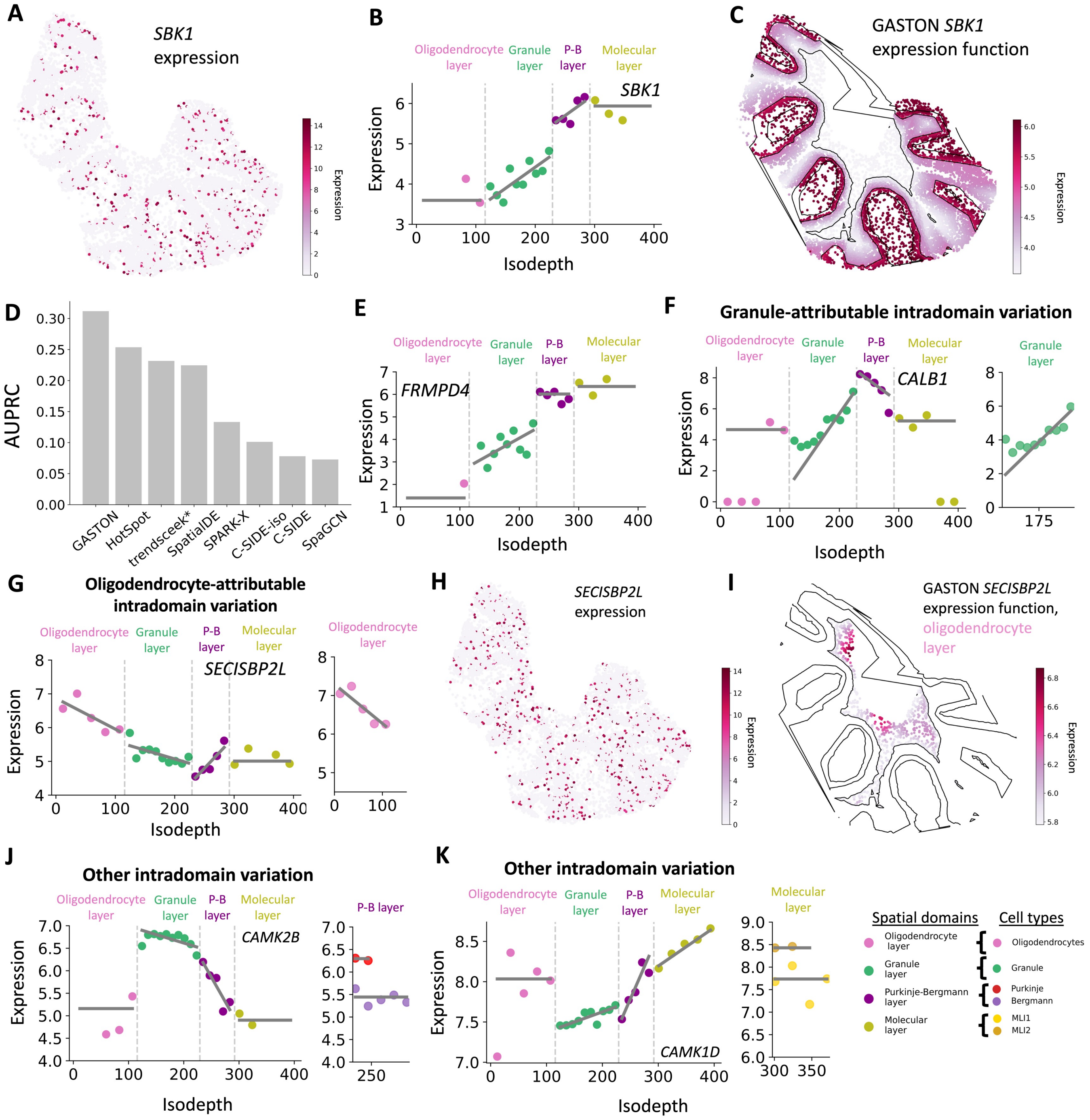


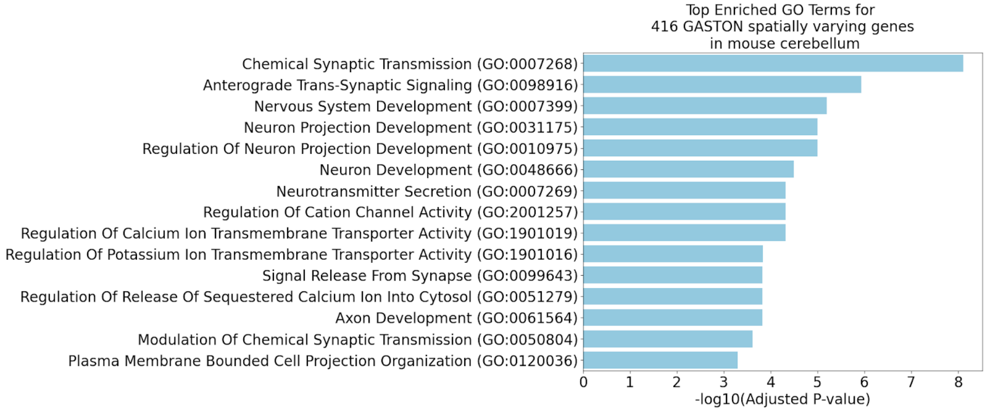


GASTON – cerebellum (spatial domains)

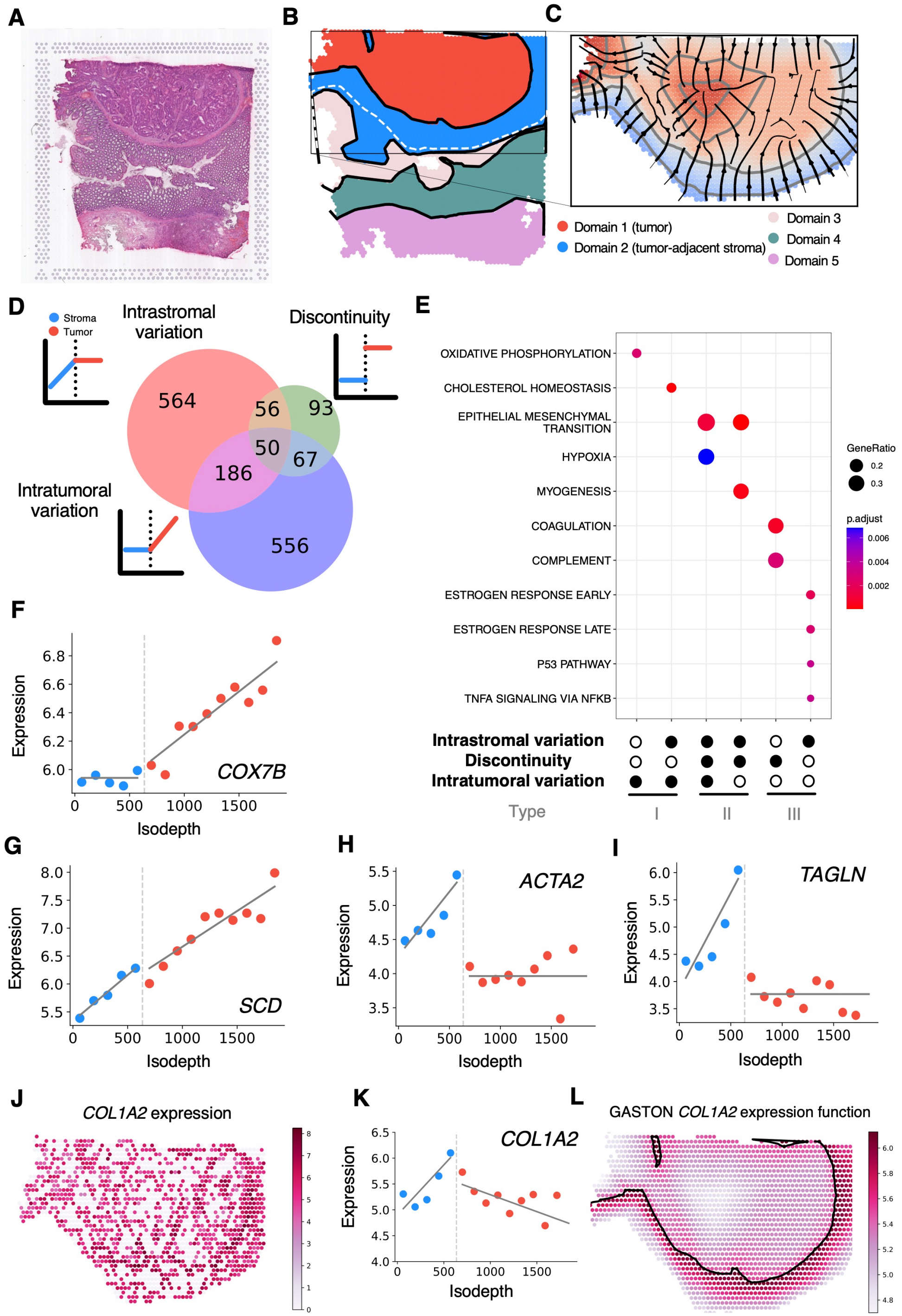


GASTON – cerebellum (spatial expression patterns)

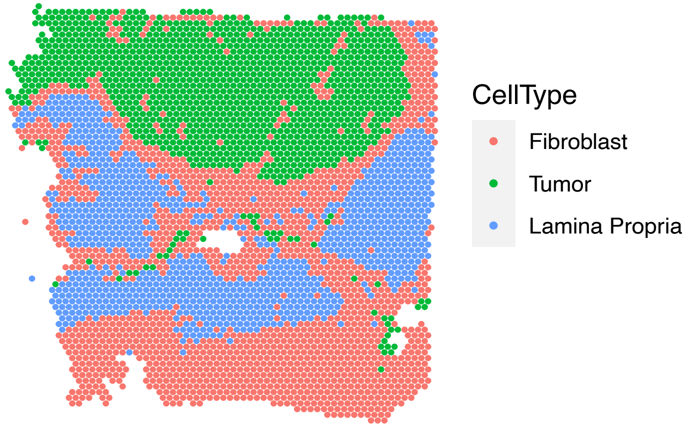




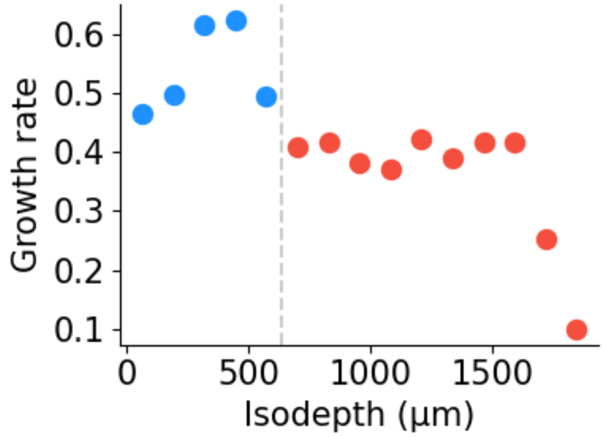
GASTON – colorectal tumor



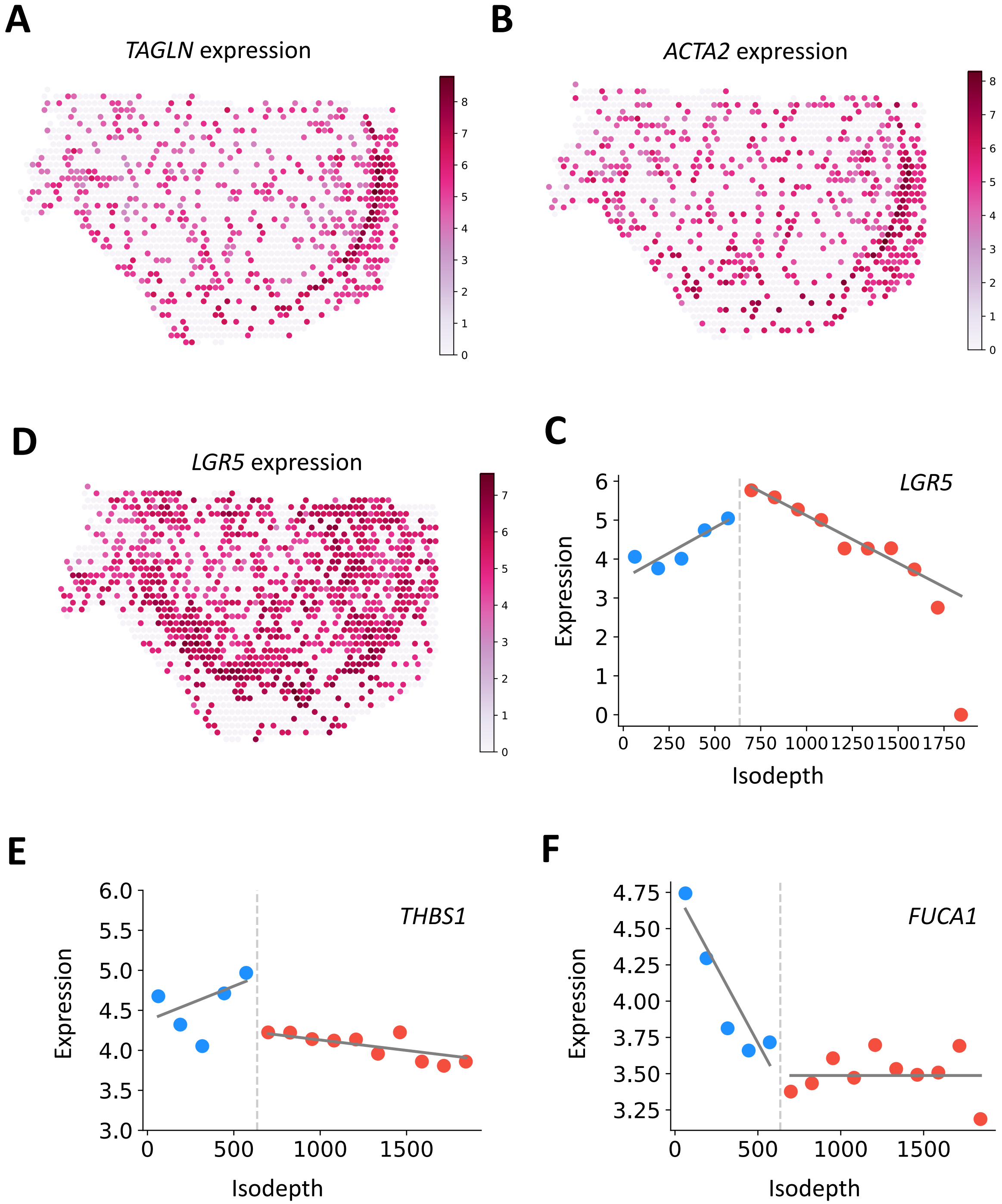
Seurat cell types



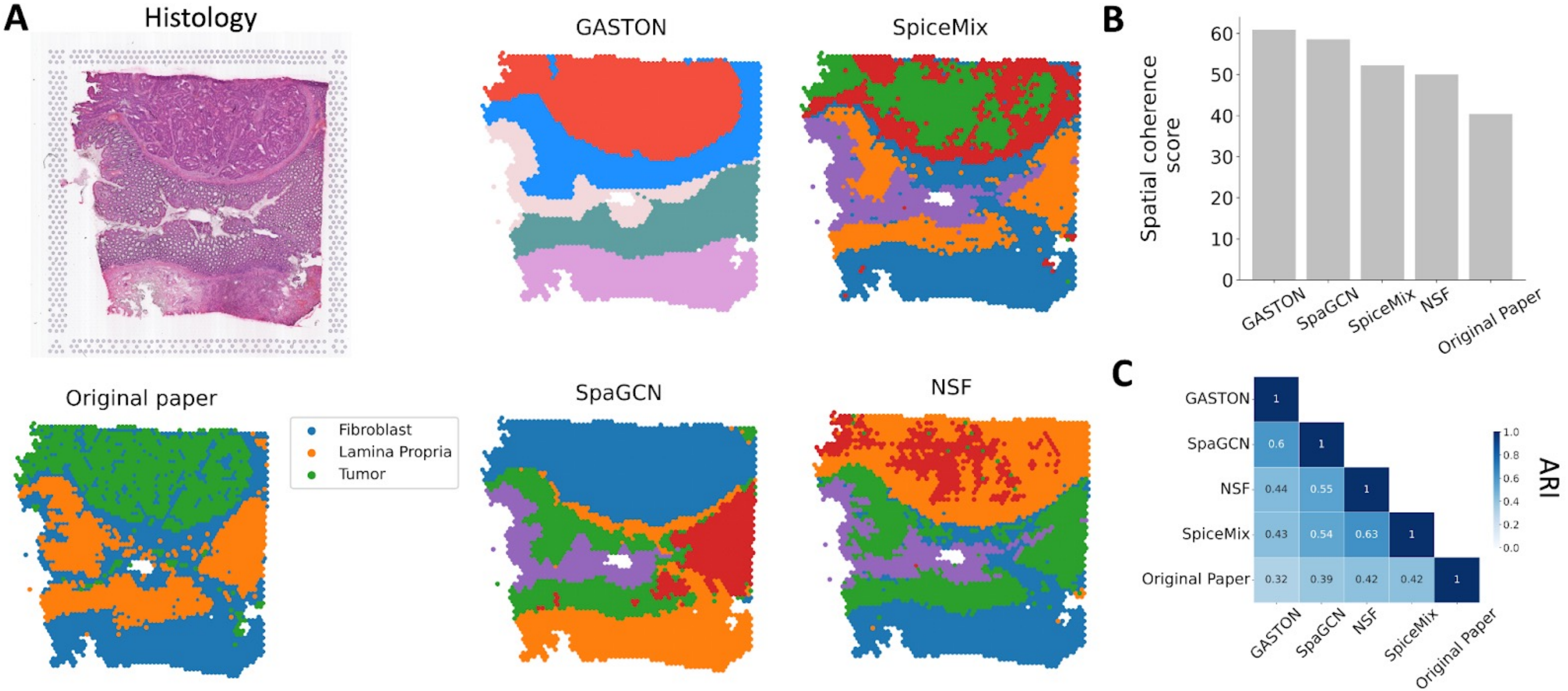
Tumor growth rate



GASTON – colorectal tumor (more patterns)



Comparison of domains on colorectal tumor

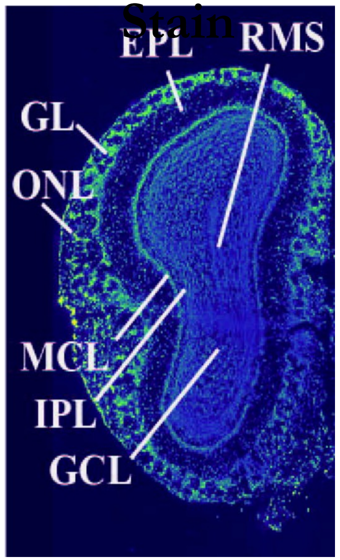


Olfactory bulb (Stereo-seq)

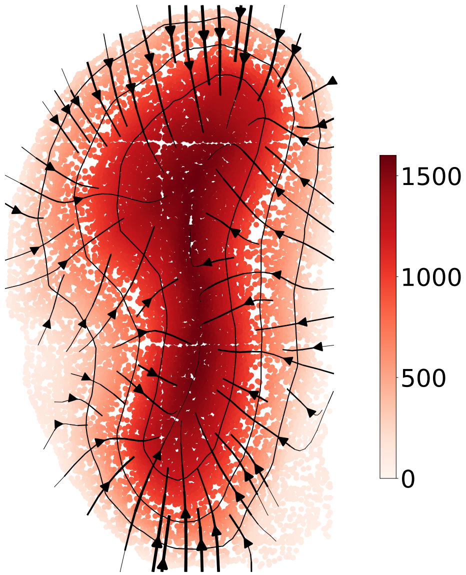
9,825 spots × 27,106 genes

- ● Olfactory nerve layer (ONL)
- ● Glomerular layer (GL)
- ● External plexiform layer (EPL)
- ● Mitral cell layer (MCL)
- ● Internal plexiform layer (IPL)
- ● Granule cell layer (GCL)
- ● Rostral migratory stream (RMS)

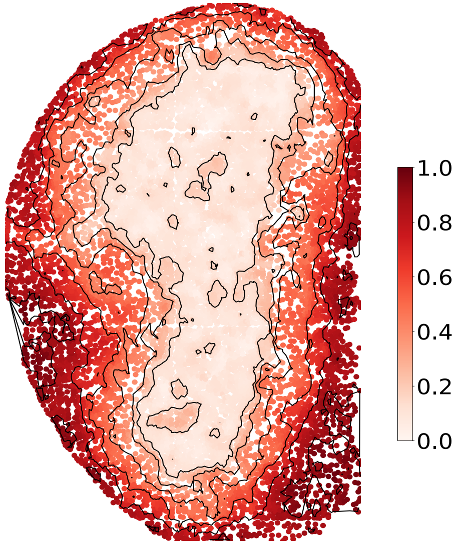
DAPI Stain



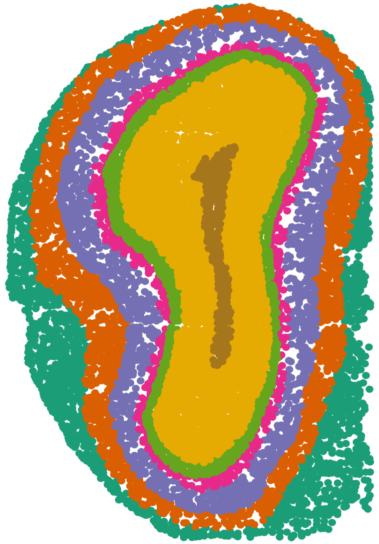
Isodepth and spatial gradients



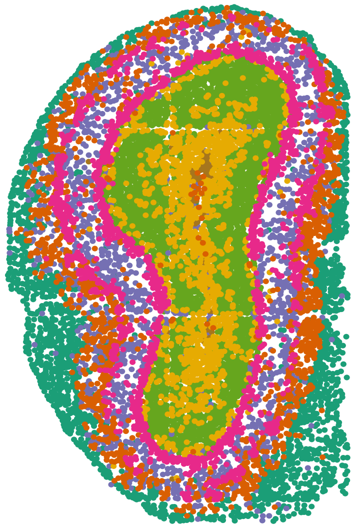
SpaceFlow (diffusion pseudotime)

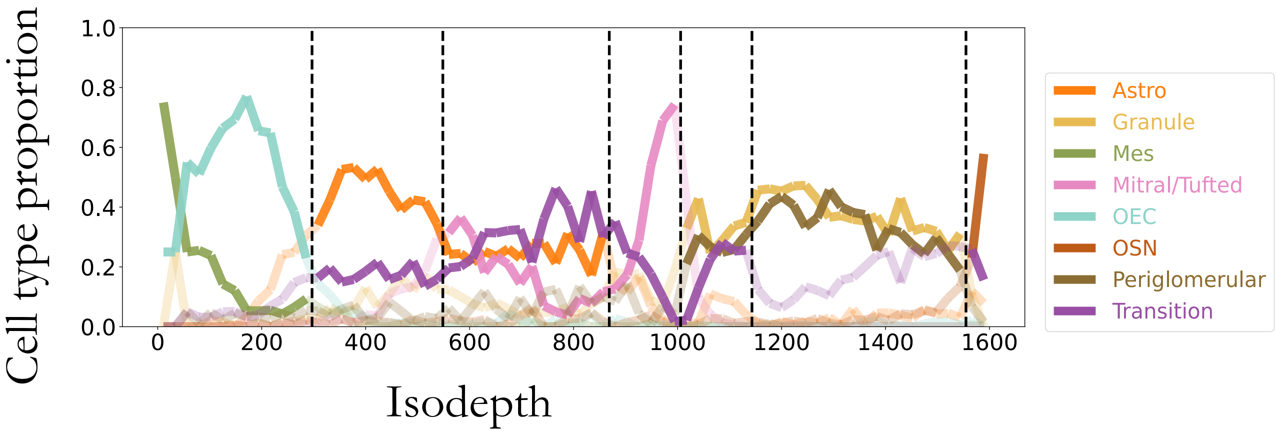


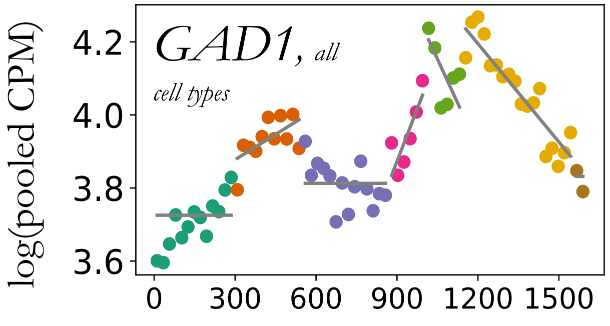
GASTON



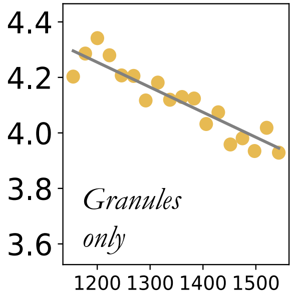
SpaGCN



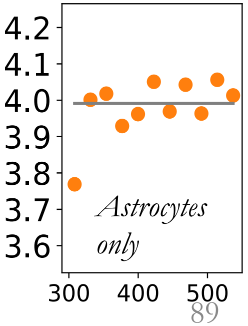




Cell type-*attributable* gradient

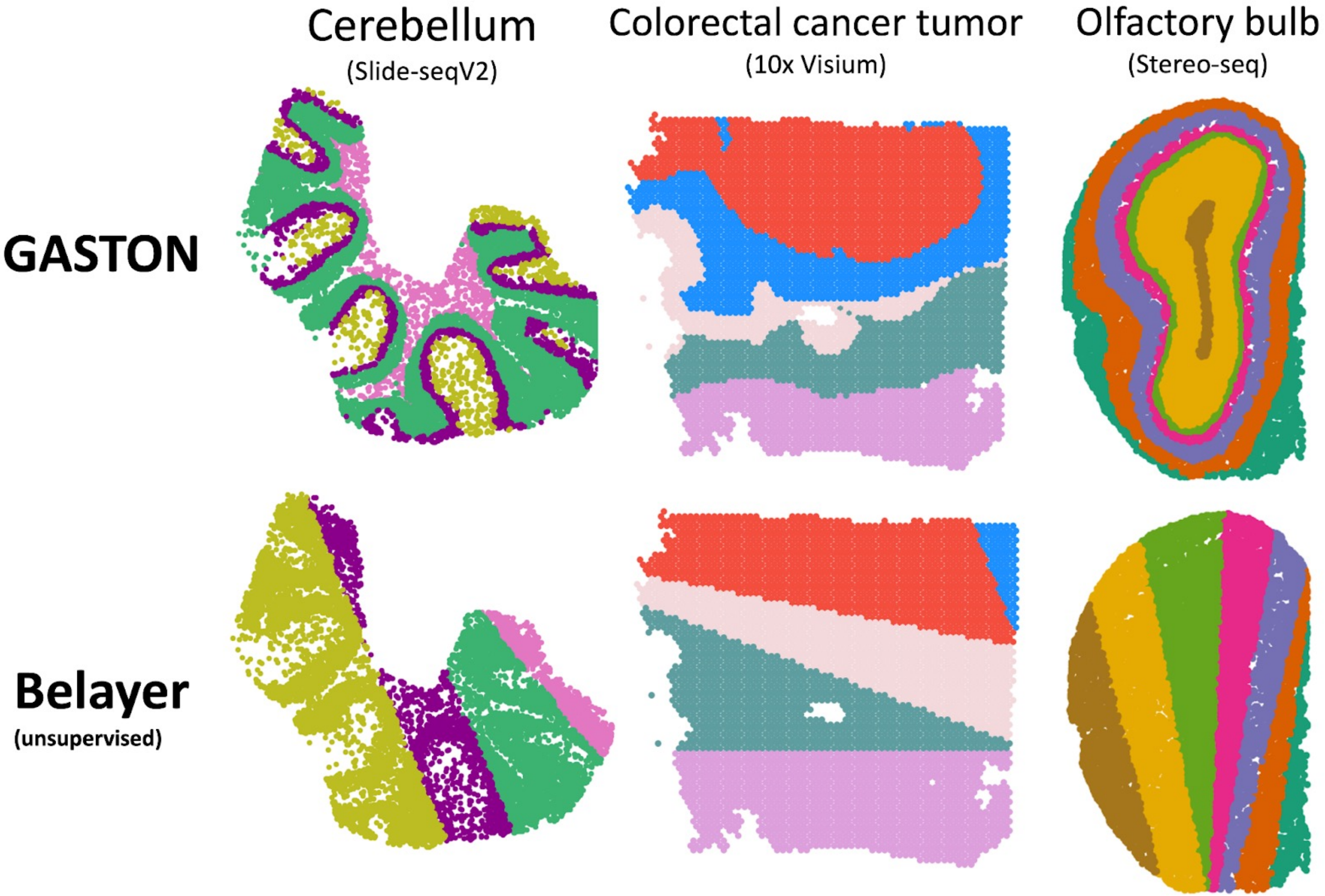


Other attributable gradient

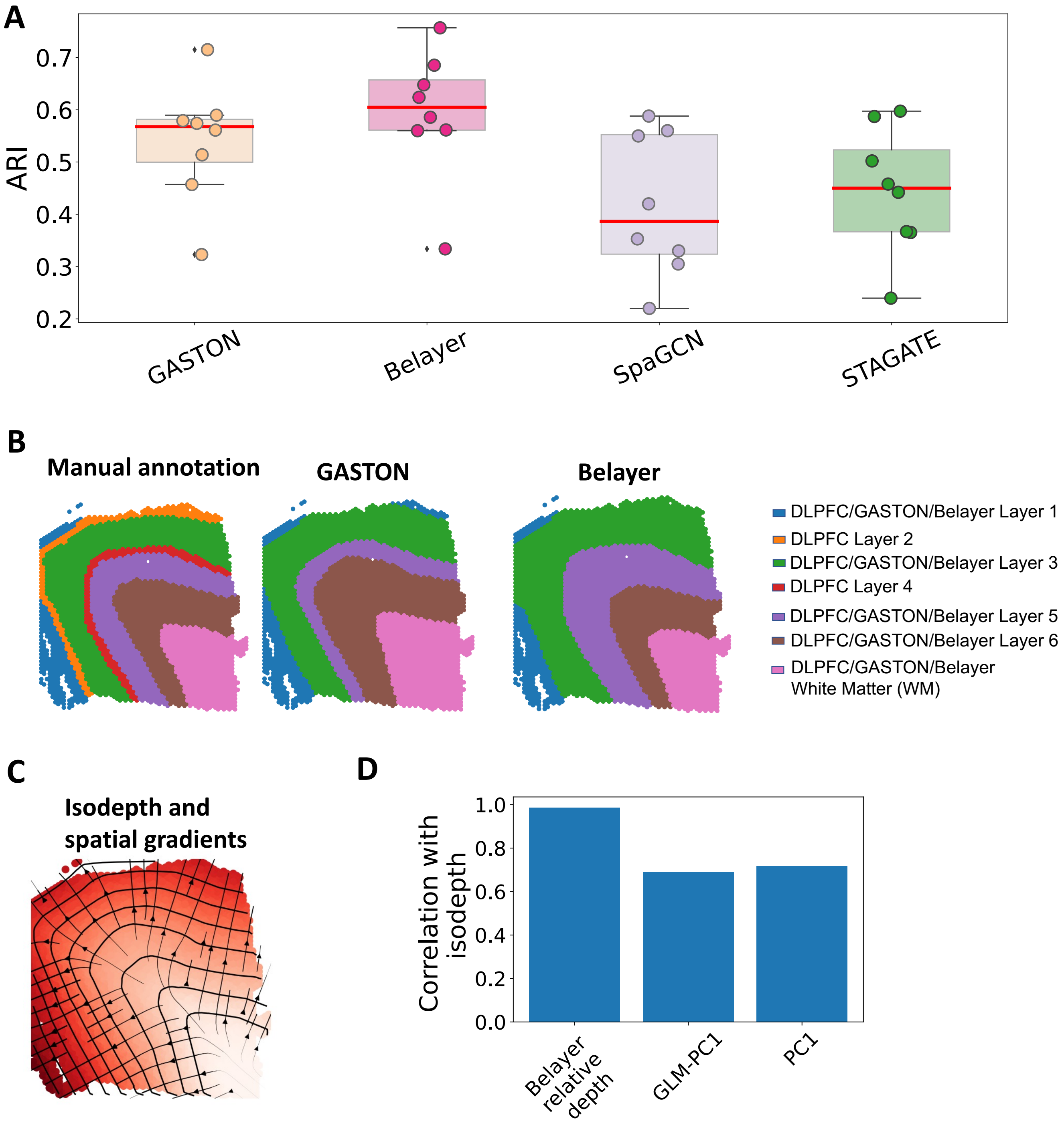


Isodepth

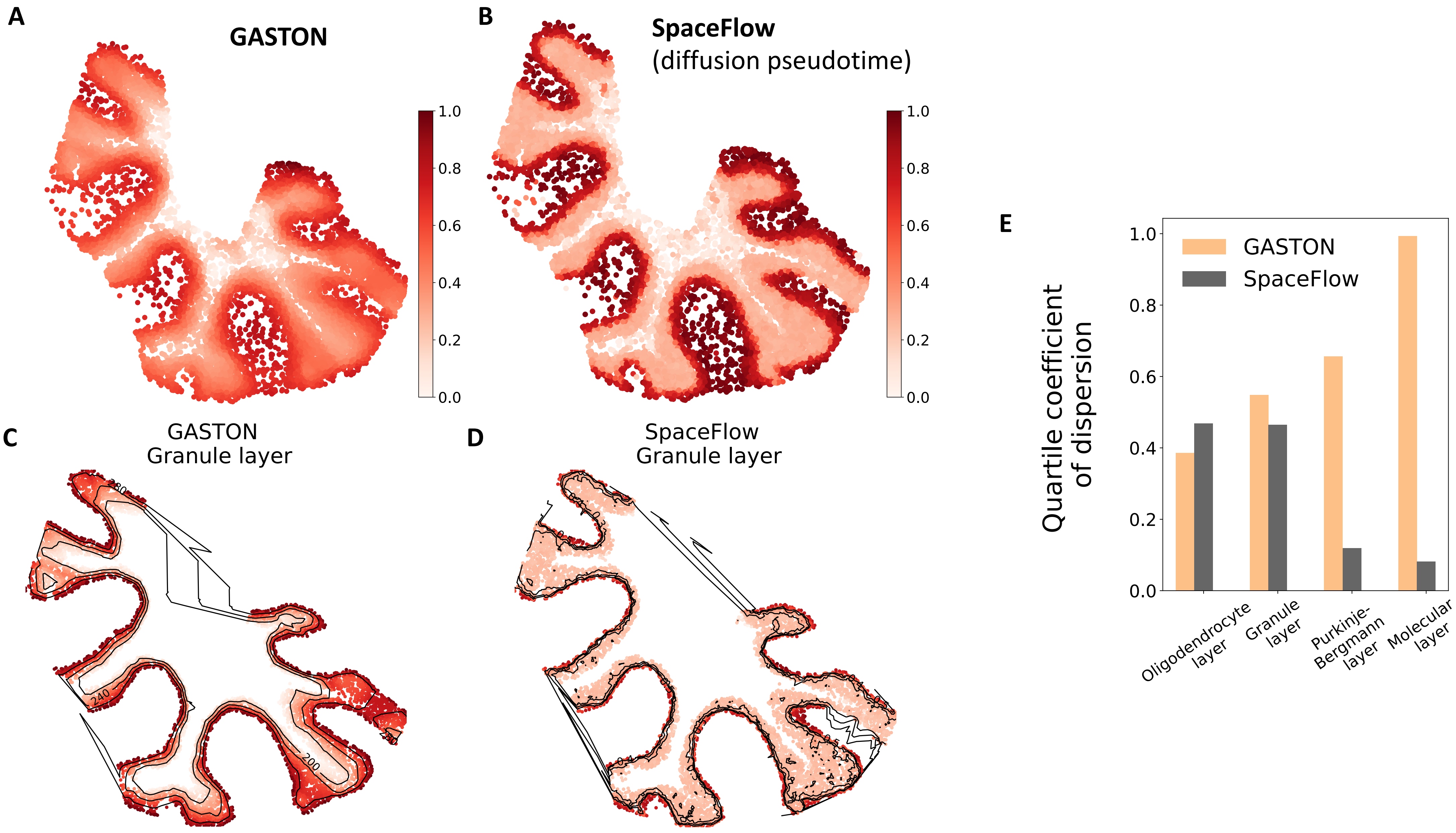
Comparison b/w GASTON and Belayer



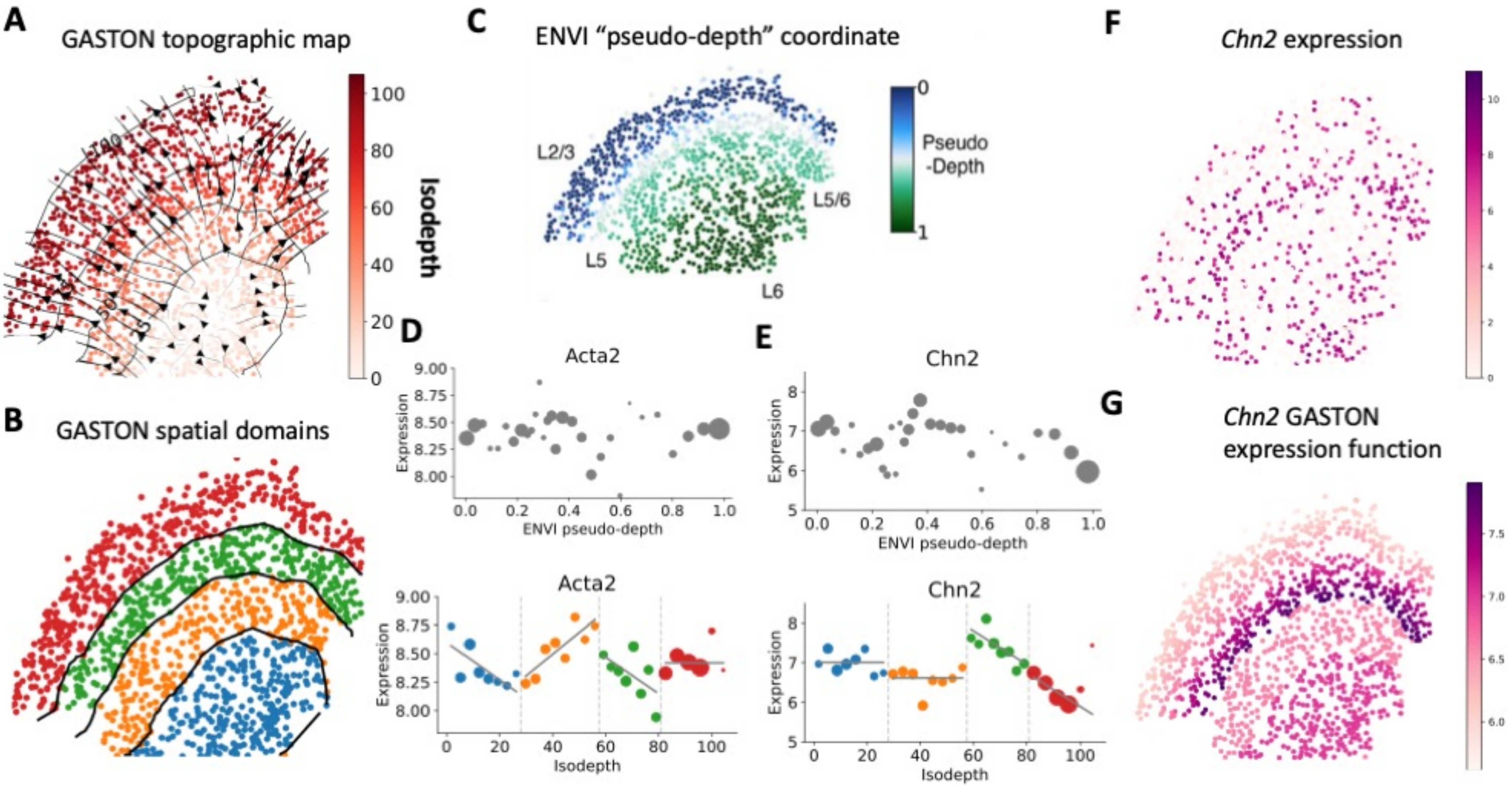
GASTON – DLPFC



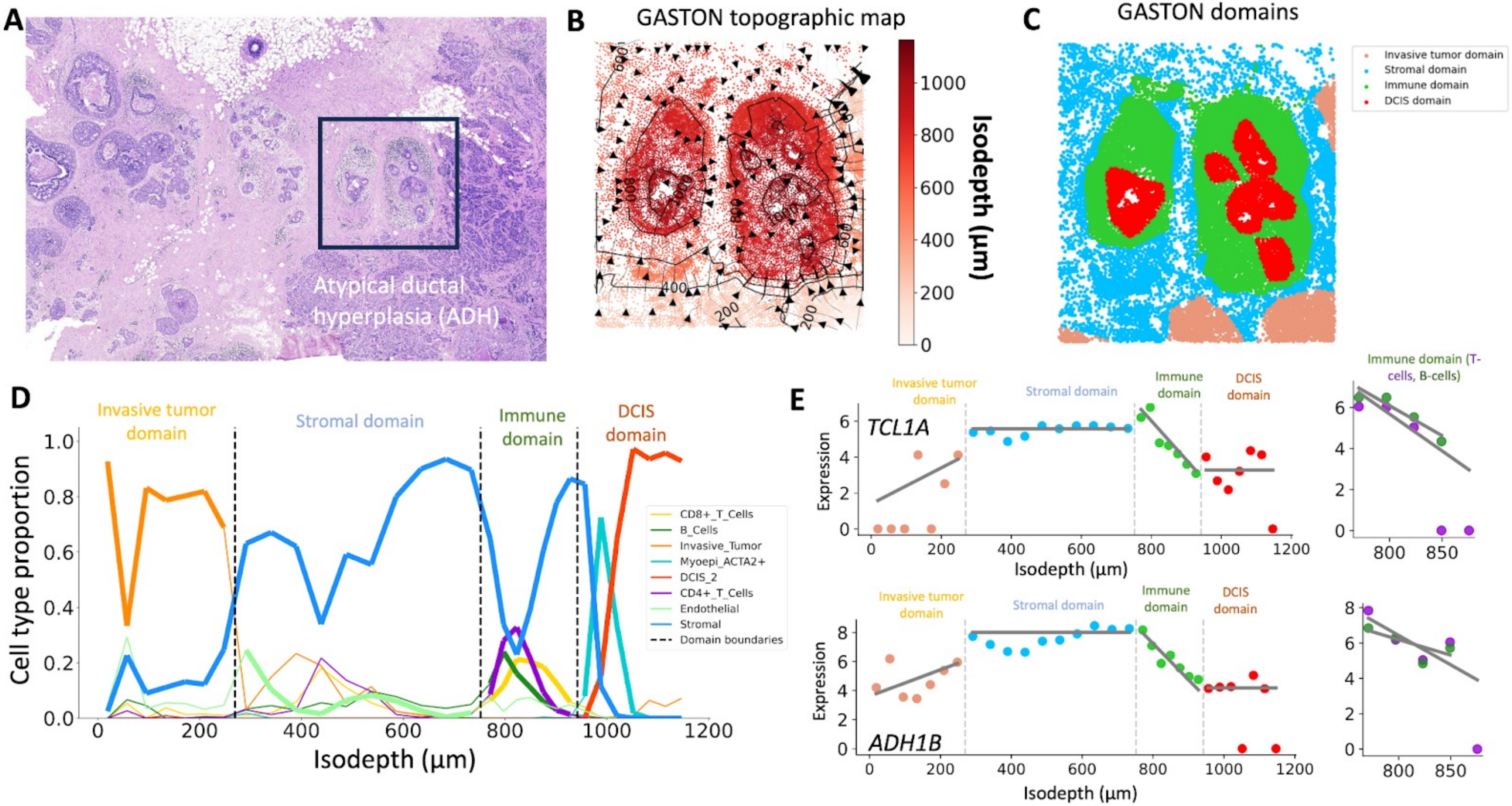
GASTON – SpaceFlow comparison (cerebellum)

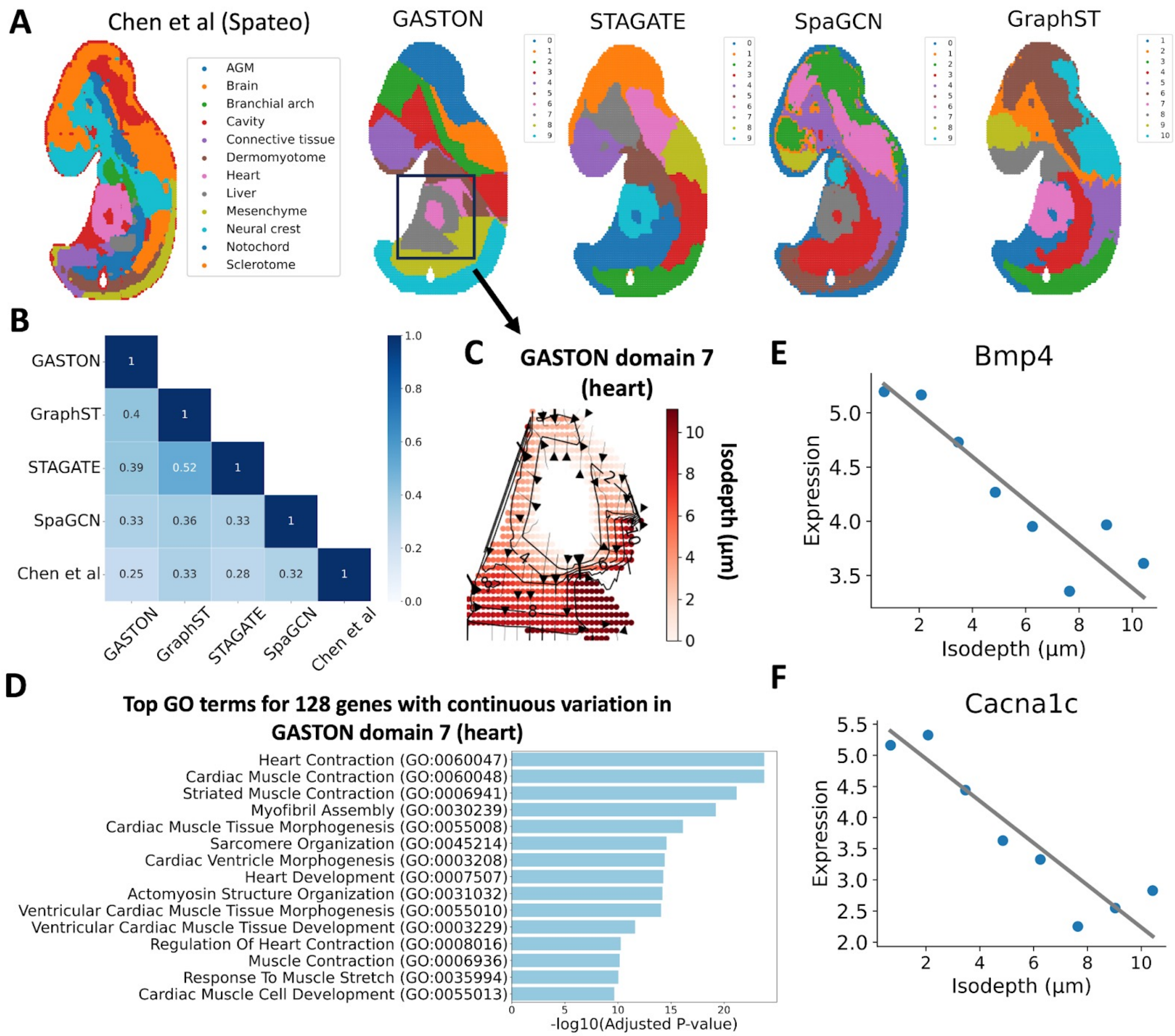


GASTON – mouse primary motor cortex (MERFISH)



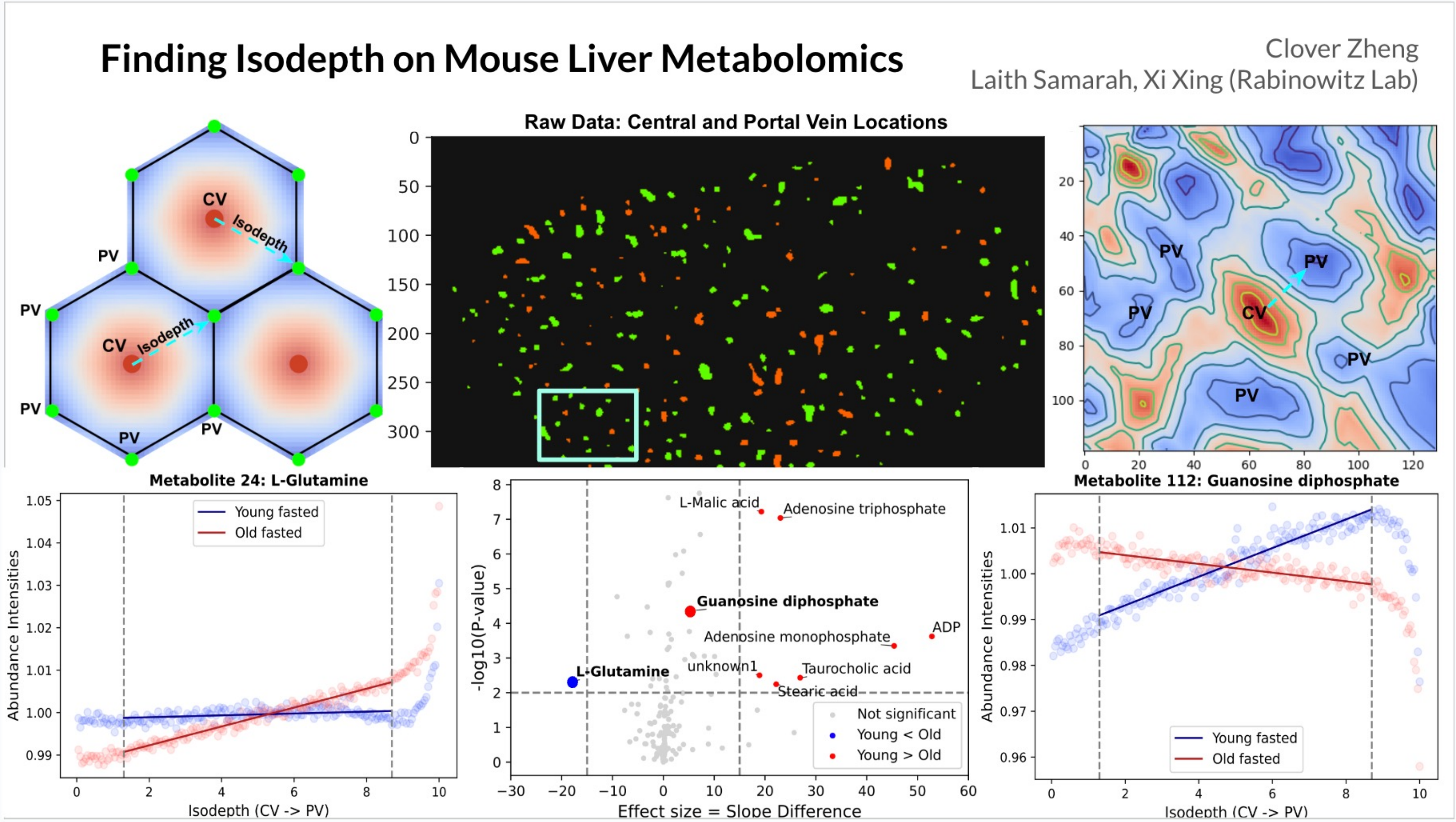
GASTON – breast cancer (10x Xenium)





GASTON – mouse embryo day 9.5 (Stereo-seq)

Application of GASTON to metabolomics (Clover Zheng)

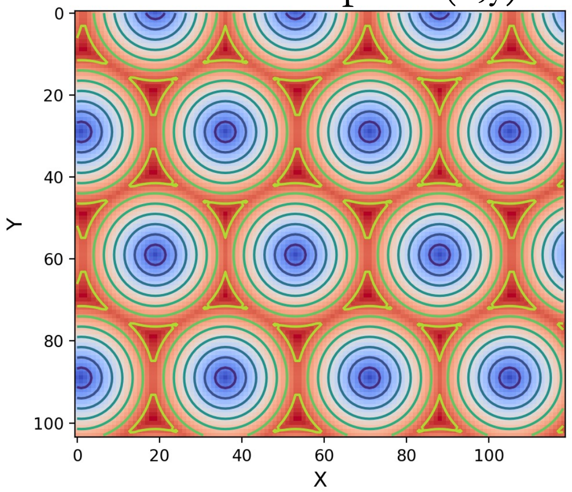


Finding Isodepth on Mouse Liver Metabolomics

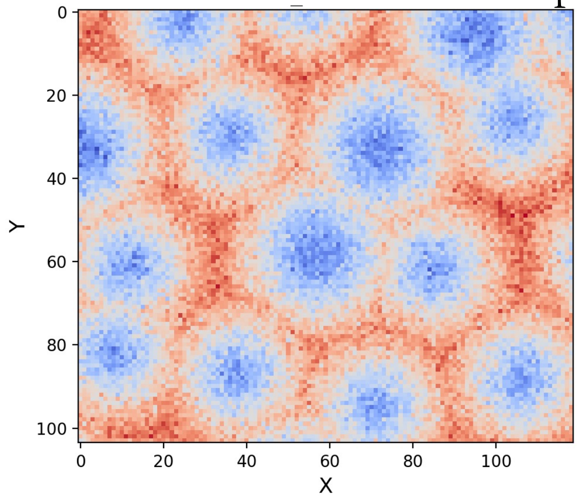
Clover Zheng
Laith Samarah, Xi Xing (Rabinowitz Lab)

Testing GASTON w/ simulated hexagonal geometries:

True isodepth d(x,y)



GASTON-estimated isodepth



GMM yields less biased estimate of altered subnetwork size

MLE: $$\widehat{A}_{\text{MLE}} = \underset{\substack{S \subseteq V \\ S \text{ connected}}}{\operatorname{argmax}} \left(\frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v \right)$$

vs

GMM: Fit vertex scores X_v to GMM

$$X_v \sim (1-\alpha) \cdot N(0,1) + \alpha \cdot N(\mu, 1)$$

and estimate GMM parameters $\widehat{\alpha}_{\text{GMM}}, \widehat{\mu}_{\text{GMM}}$

We prove (ICML 2021): GMM yields asymptotically unbiased estimates of α, μ, i.e.

$$\lim_{n\to\infty} |\widehat{\alpha}_{\text{GMM}} - \alpha| = 0$$

$$\lim_{n\to\infty} |\widehat{\mu}_{\text{GMM}} - \mu| = 0$$

Model mis-specification helps!
(Fitting ASD with GMM)

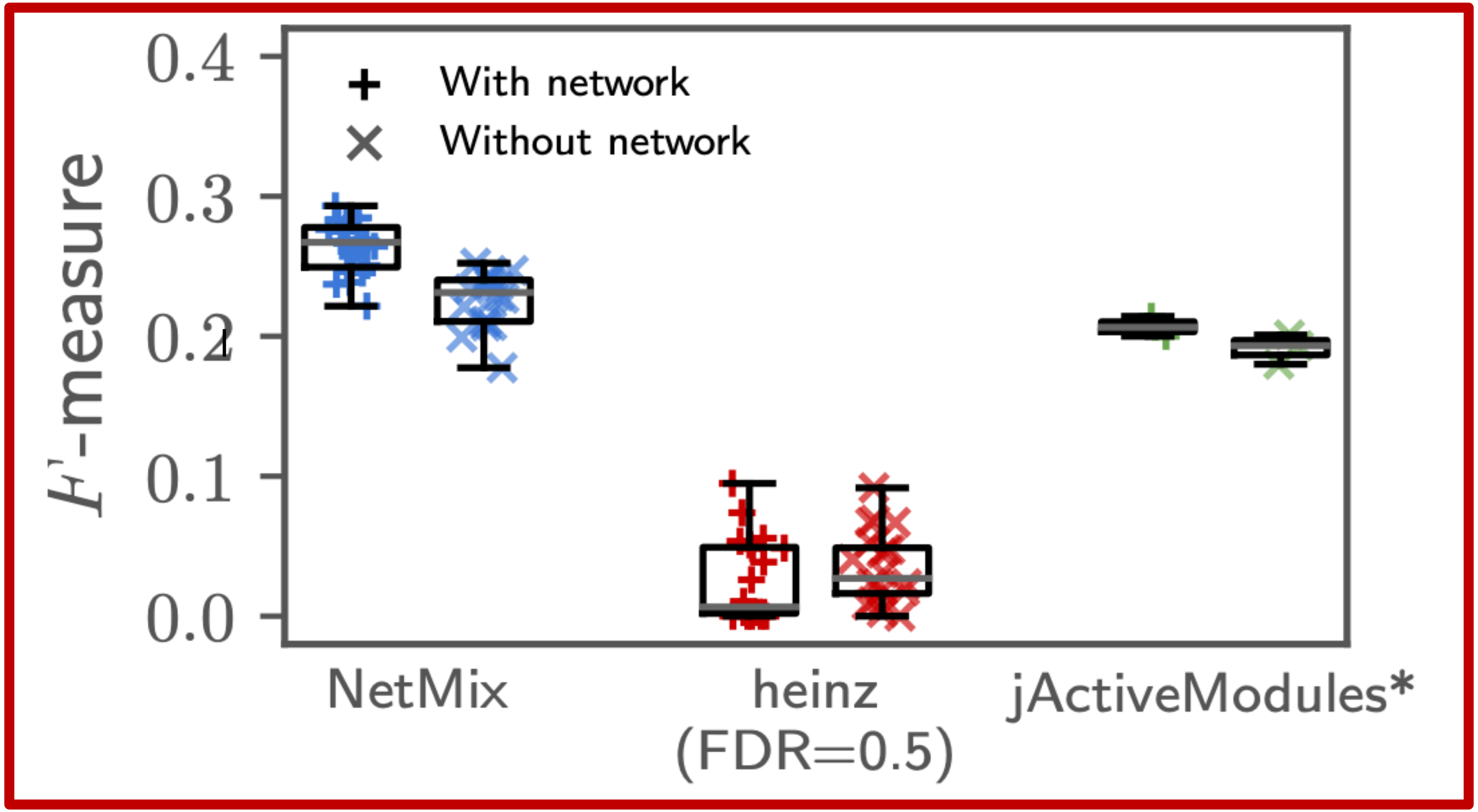
α = proportion of vertices in altered subnetwork

μ = mean of altered subnetwork distribution

Challenge: Connectivity is a weak topological constraint!

Networks have small diameter – most subnetworks are "almost connected"

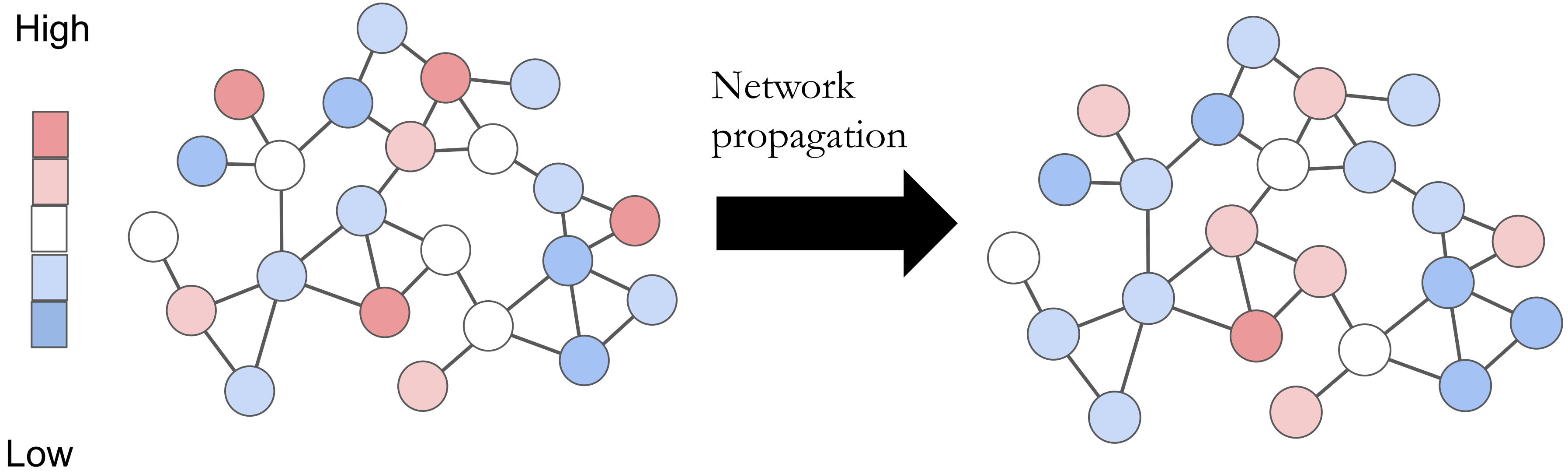
Algorithms not much better compared to not using interaction network

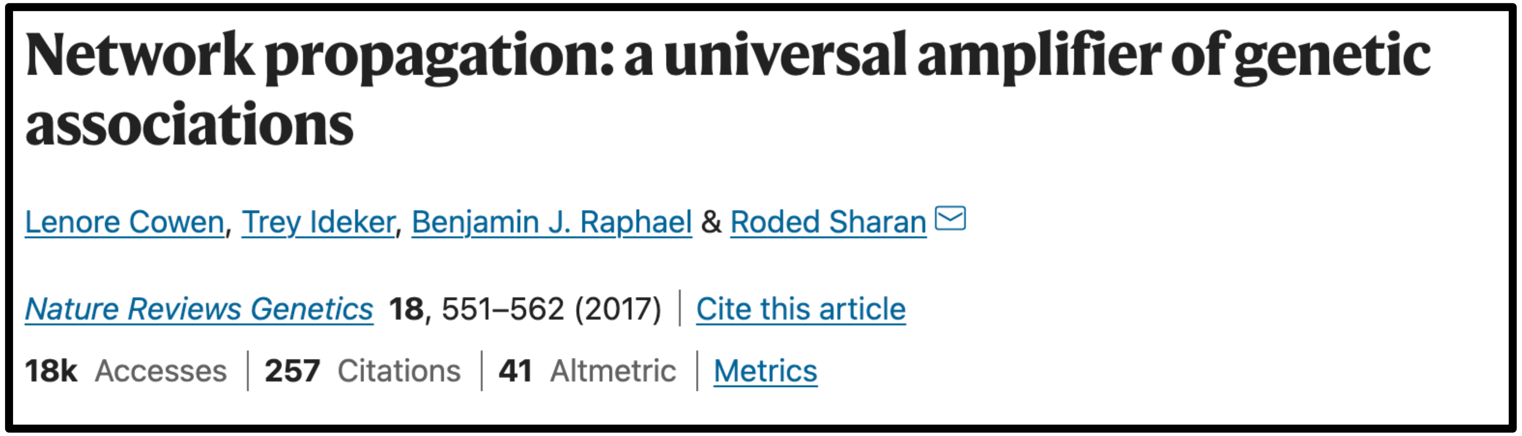


Simulations from our generative model where altered subnetwork is **connected subgraph**

Network propagation (network diffusion)

Use of random walks to “propagate”/smooth vertex scores across network



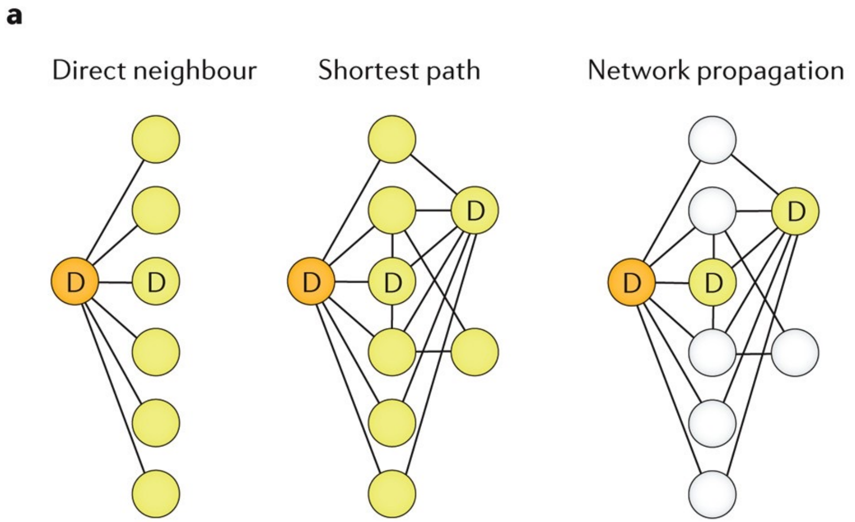


Network propagation: a universal amplifier of genetic associations

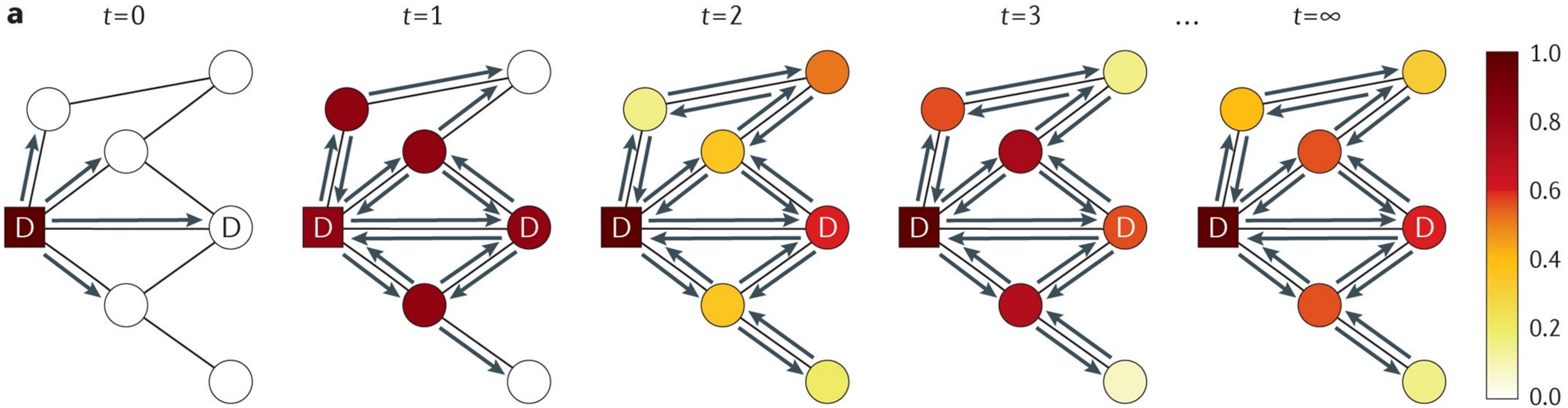
Lenore Cowen, Trey Ideker, Benjamin J. Raphael & Roded Sharan

Nature Reviews Genetics **18**, 551–562 (2017) | Cite this article

18k Accesses | **257** Citations | **41** Altmetric | Metrics

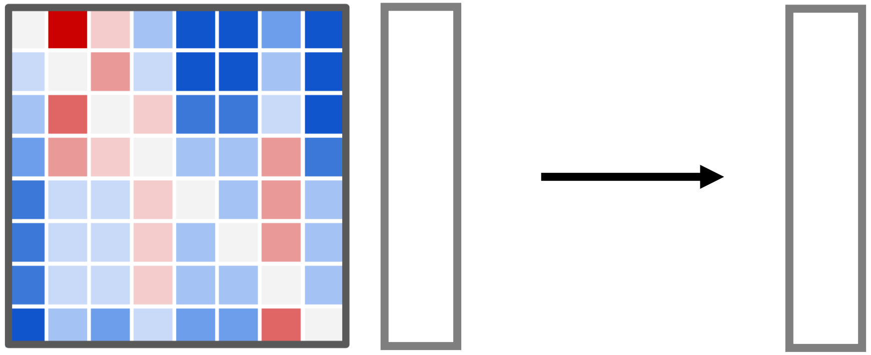


Network propagation uses global network structure



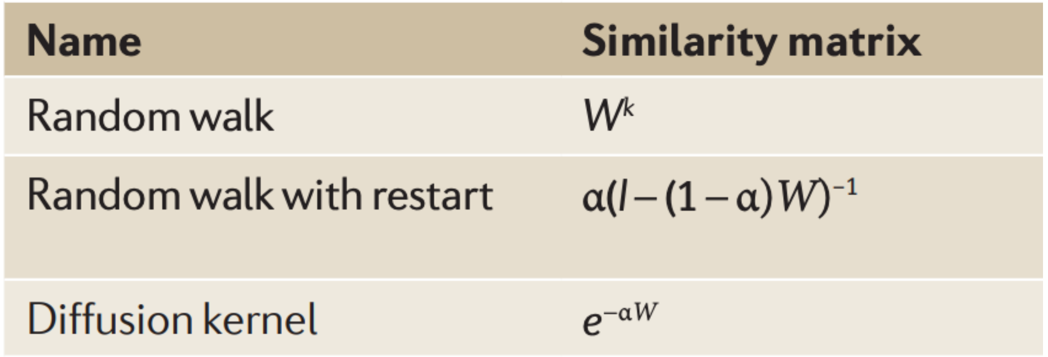
Cowen et al (Nature Reviews Genetics 2017)

Network propagation = Matrix-vector multiplication



Random walk similarity matrix

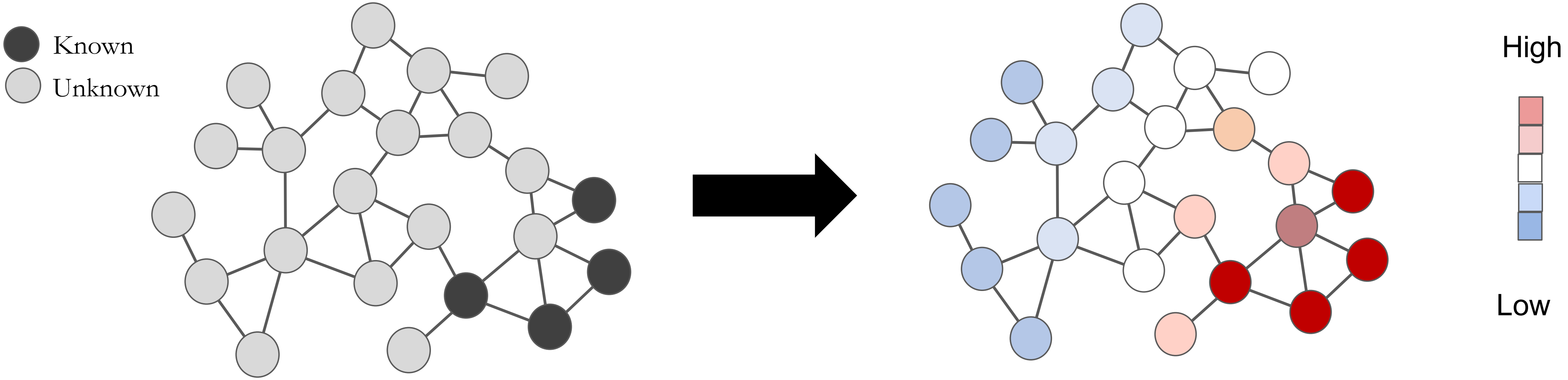
Vertex scores



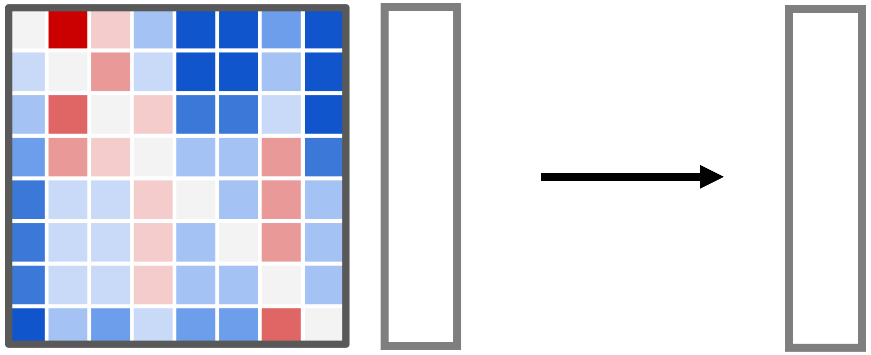
Name	Similarity matrix
Random walk	W^k
Random walk with restart	$\alpha(I-(1-\alpha)W)^{-1}$
Diffusion kernel	$e^{-\alpha W}$

Cowen et al (Nature Reviews Genetics 2017)

Network propagation is standard for ranking vertices



Rank vertices based on similarity to vertices w/ known characteristics e.g. genes associated with a specific disease (binary vertex scores X_v)



Random walk similarity matrix

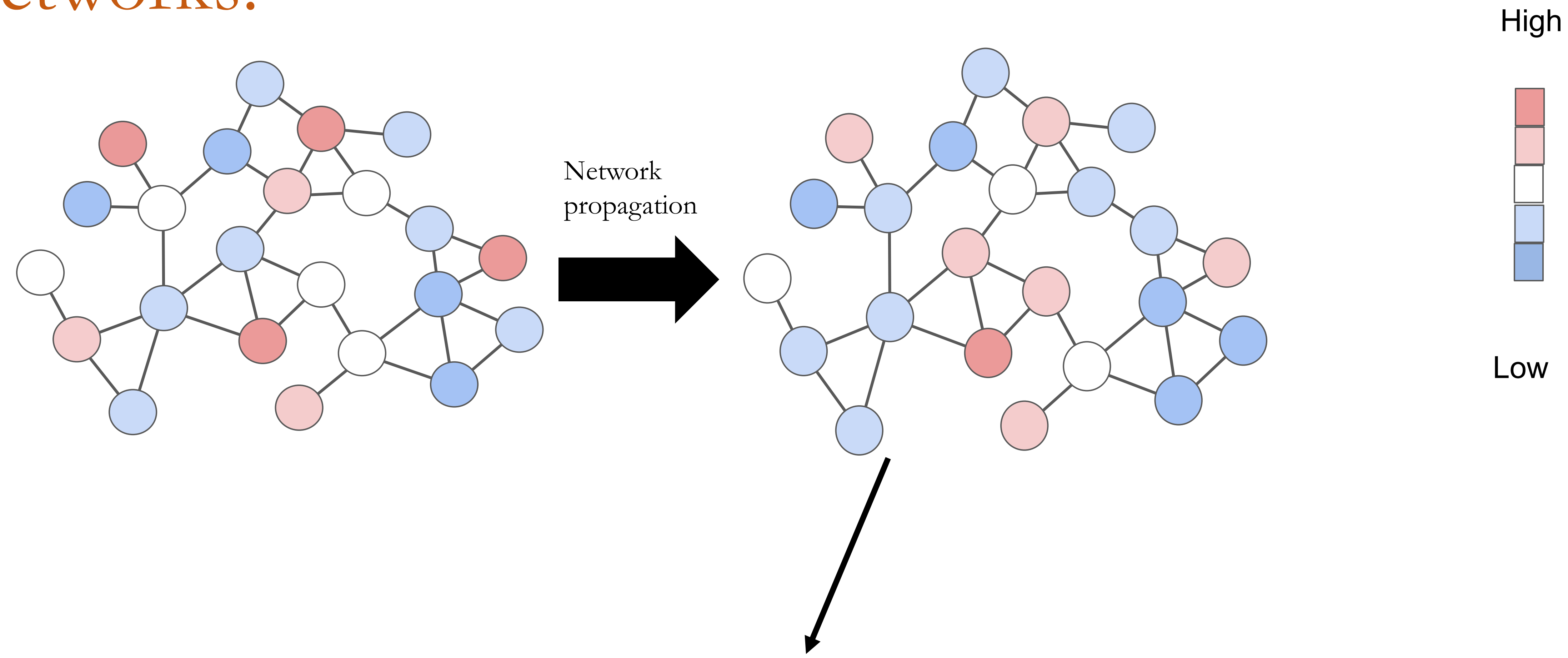
Vertex scores

Google

PageRank

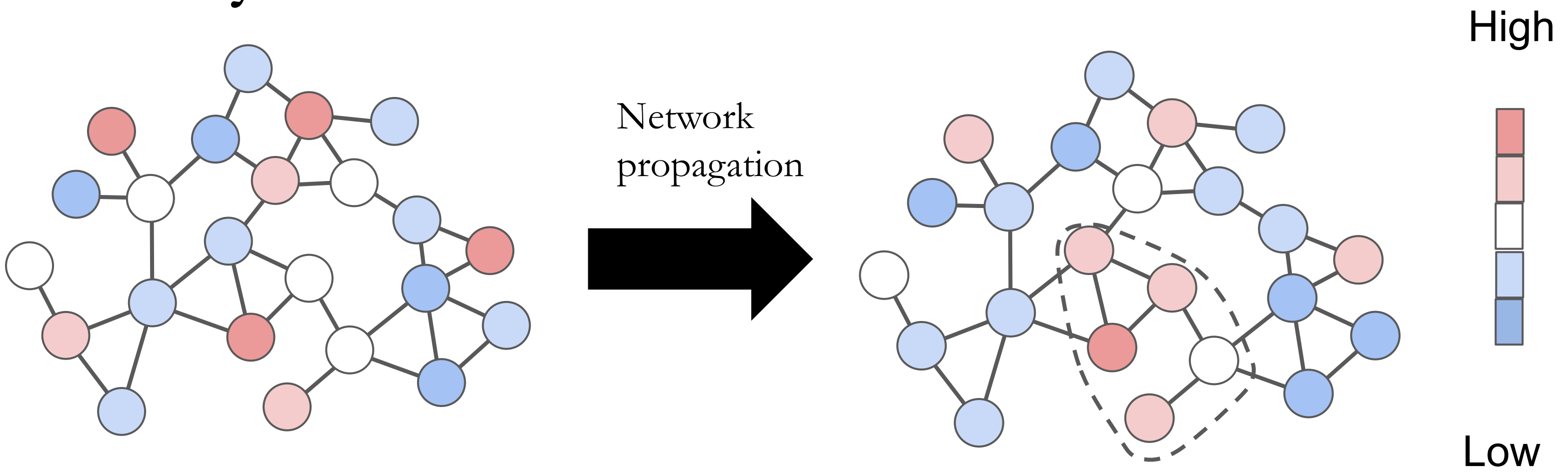
Personalized PageRank is **asymptotically optimal** for ranking in random graph models (PNAS 2017)

How to use network propagation to identify altered subnetworks?

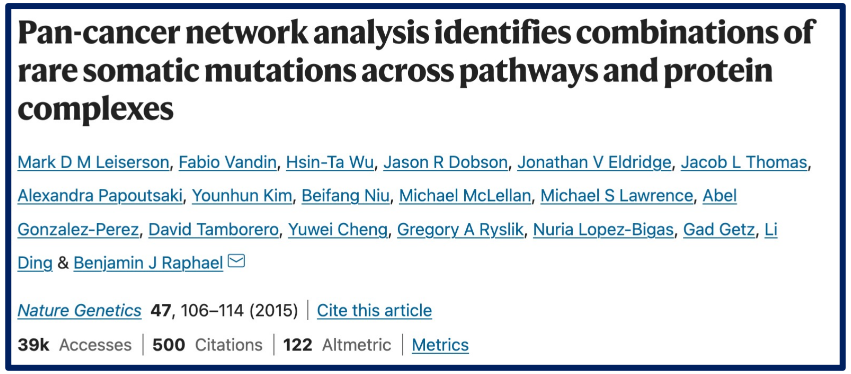


Question: how to identify altered subnetwork from propagated gene scores?

Existing network propagation methods use ad hoc heuristics to identify altered subnetworks



HotNet2



Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes

Mark D M Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding & Benjamin J Raphael

Nature Genetics **47**, 106–114 (2015) | Cite this article

39k Accesses | **500** Citations | **122** Altmetric | Metrics

PRINCE

Associating Genes and Protein Complexes with Disease via Network Propagation

Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, Roded Sharan

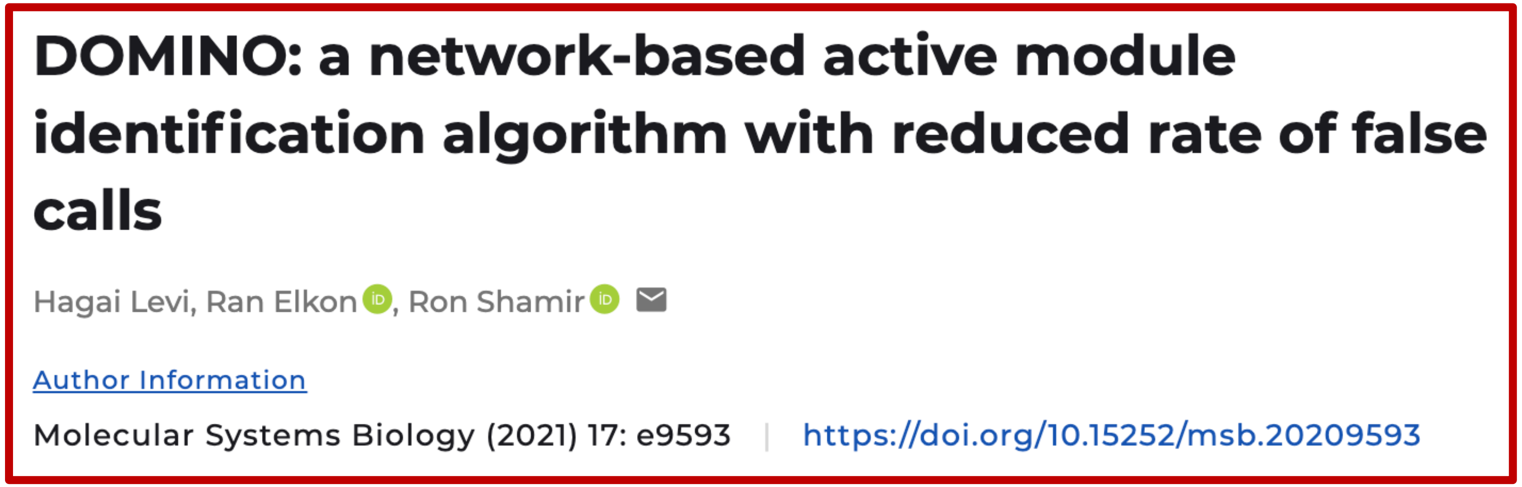
Published: January 15, 2010 • https://doi.org/10.1371/journal.pcbi.1000641

Ex: **PRINCE**: "We aim at inferring densely connected protein complexes that contain high scoring proteins … we start with the top 100 [propagated] scoring proteins as seeds … To each seed we iteratively add a neighboring protein with the highest score … A refinement phase takes place where proteins are removed from a putative complex to ensure that … its member proteins are densely interacting."

Issue: These algorithms lack **rigorous statistical guarantees** – hard to investigate fundamental issues like bias

Recent work shows existing approaches biased towards "high centrality" vertices

Algorithms benchmark against existing network algorithms – can hide biases shared across methods



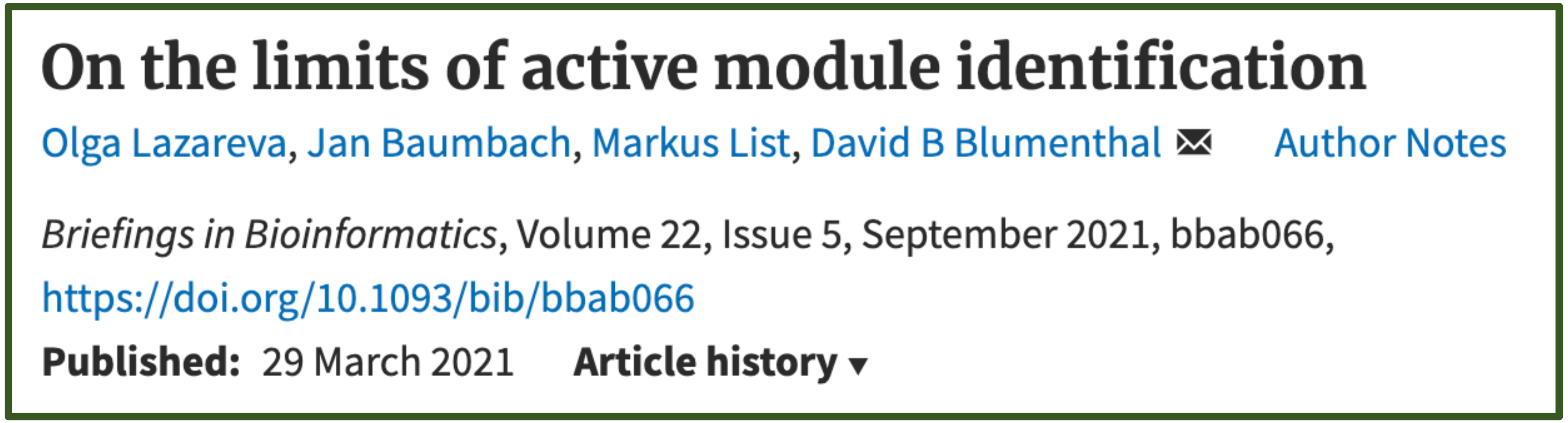
DOMINO: a network-based active module identification algorithm with reduced rate of false calls

Hagai Levi, Ran Elkon, Ron Shamir

Author Information

Molecular Systems Biology (2021) 17: e9593 | https://doi.org/10.15252/msb.20209593

*"Our study reports on a different bias that is prevalent in AMI solutions: **their tendency to report non-specific GO terms**. …we observed that many enriched GO terms also appear on permuted datasets, suggesting that such enrichment stems from some proprieties of the network, algorithm, or the data that bias the results."*



On the limits of active module identification

Olga Lazareva, Jan Baumbach, Markus List, David B Blumenthal Author Notes

Briefings in Bioinformatics, Volume 22, Issue 5, September 2021, bbab066, https://doi.org/10.1093/bib/bbab066

Published: 29 March 2021 **Article history** ▾

*"Our results indicate that classical but also supposedly bias-aware [altered subnetwork algorithms] extract disease modules **based on the node degree**"*

Our work:

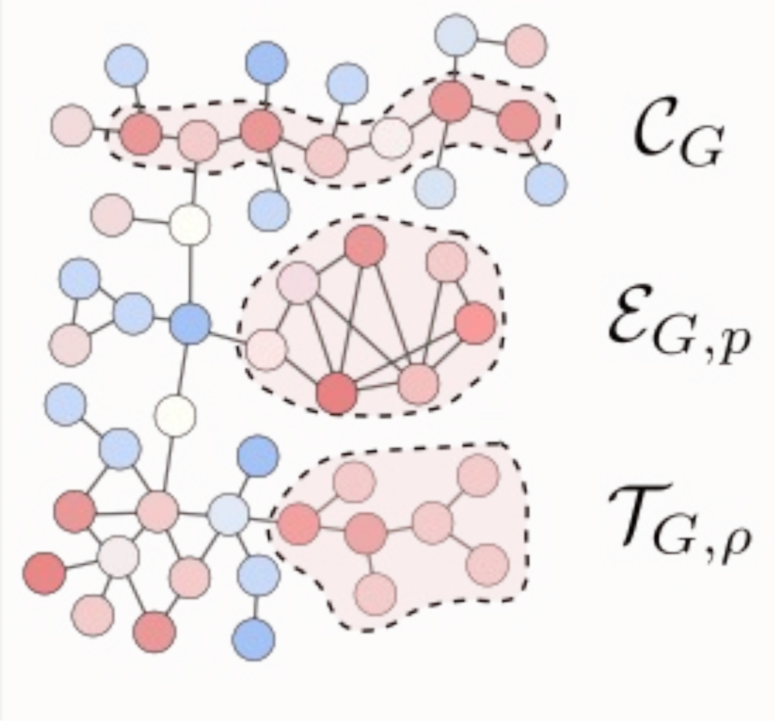
- Extend altered subnetwork generative model
  - Model different altered subnetwork topologies ("**subnetwork families**")
  - Derive <u>propagation family</u> – "approximates" subnetworks found by network propagation

- **<u>NetMix2</u>** algorithm for altered subnetwork identification with different subnetwork families
  - w/ propagation family: principled network propagation algorithm for altered subnetwork identification

- Simple baselines for evaluating network algorithms – *"scores only"* and *"network only"*

Generative model: Altered Subnetwork Distribution

- G=(V, E) is interaction network
- $\mathcal{S}$ is **subnetwork family** (set of subsets of V)
- $A \in \mathcal{S}$ is the altered subnetwork



Vertex scores $(X_v)_{v \in V}$ are distributed as

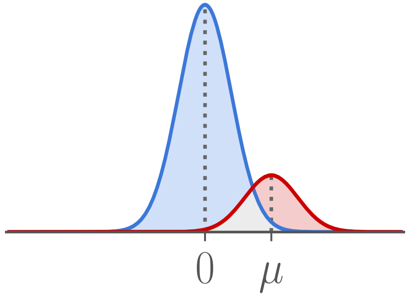
$$X_v \sim \begin{cases} \mathcal{D}_\mathrm{a}, & \text{if } v \in A, \\ \mathcal{D}_\mathrm{b}, & \text{otherwise} \end{cases}$$

$\mathcal{D}_\mathrm{a}$ = altered distribution (unknown)

$\mathcal{D}_\mathrm{b}$ = background distribution (typically known)

Example of distributions: z-scores

$$\mathcal{D}_\mathrm{a} = N(\mu, 1)$$
$$\mathcal{D}_\mathrm{b} = N(0, 1)$$



Examples of subnetwork families:

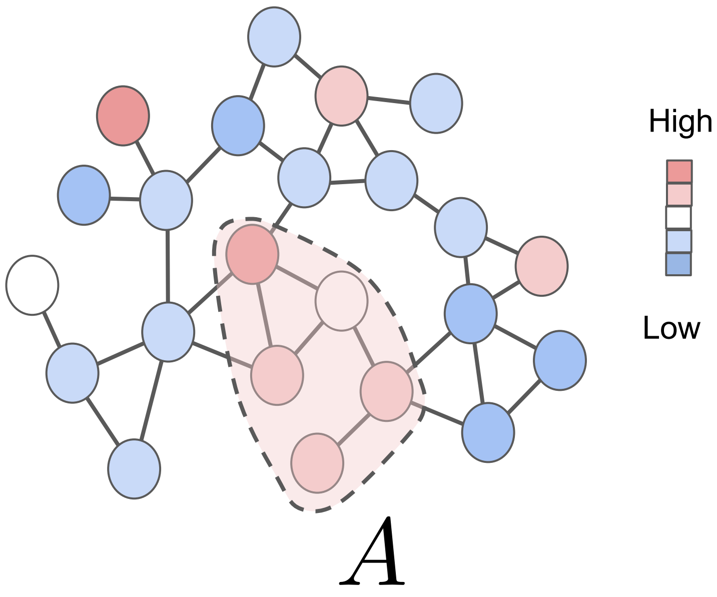
Connected family $\mathcal{S} = \mathcal{C}_G$ = connected subgraphs S ⊆ V

Edge-dense family $\mathcal{S} = \mathcal{E}_{G,p}$ = subgraphs with density(S) > p

Cut family $\mathcal{S} = \mathcal{T}_{G,\rho}$ = subgraphs with cut(S) < ϱ

Generative model: Altered Subnetwork Distribution

- G=(V, E) is interaction network
- $\mathcal{S}$ is **subnetwork family** (set of subsets of V)
- $A \in \mathcal{S}$ is the altered subnetwork

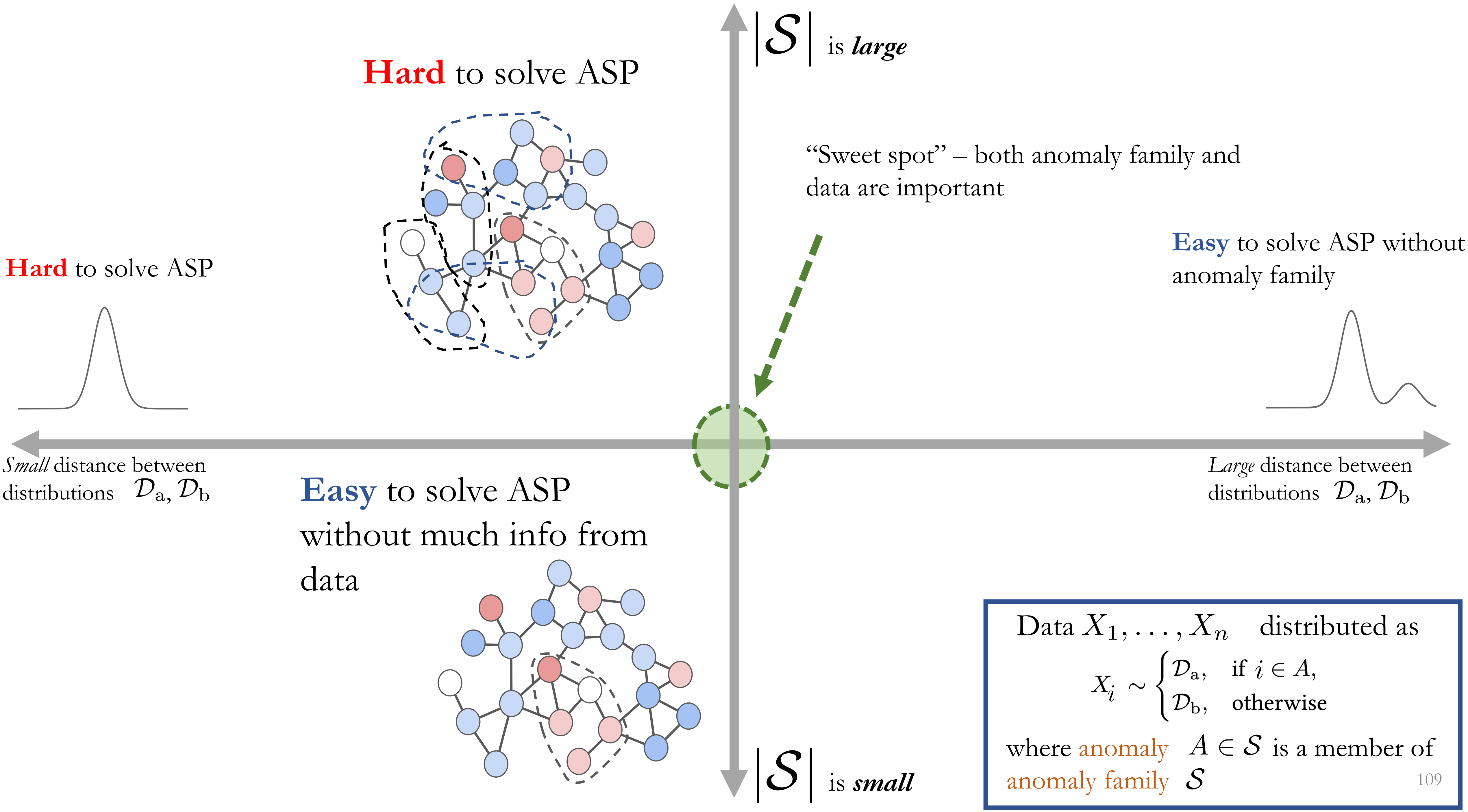


Vertex scores $(X_v)_{v \in V}$ are distributed as

$$X_v \sim \begin{cases} \mathcal{D}_\mathrm{a}, & \text{if } v \in A, \\ \mathcal{D}_\mathrm{b}, & \text{otherwise} \end{cases}$$

Altered Subnetwork Problem (ASP): Given graph G, subnetwork family $\mathcal{S}$ and vertex scores $(X_v)_{v \in V}$, find altered subnetwork A.

ASP = estimating parameters of distribution



$|\mathcal{S}|$ is ***large***

Hard to solve ASP

“Sweet spot” – both anomaly family and data are important

Hard to solve ASP

Easy to solve ASP without anomaly family

Small distance between distributions $\mathcal{D}_{\mathrm{a}}, \mathcal{D}_{\mathrm{b}}$

Large distance between distributions $\mathcal{D}_{\mathrm{a}}, \mathcal{D}_{\mathrm{b}}$

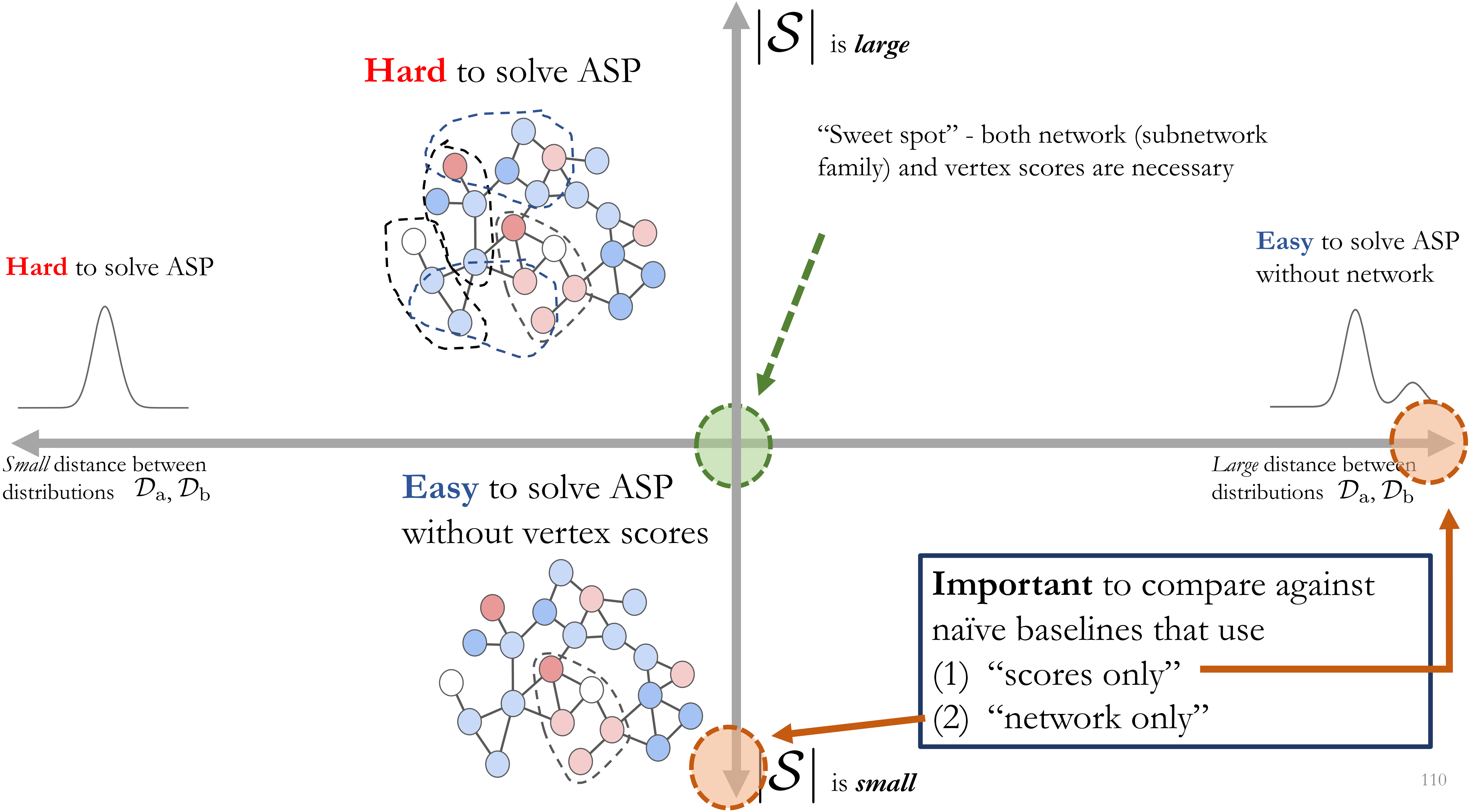
Easy to solve ASP without much info from data

$|\mathcal{S}|$ is ***small***

Data $X_1, \dots, X_n$ distributed as

$$X_i \sim \begin{cases} \mathcal{D}_{\mathrm{a}}, & \text{if } i \in A, \\ \mathcal{D}_{\mathrm{b}}, & \text{otherwise} \end{cases}$$

where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$



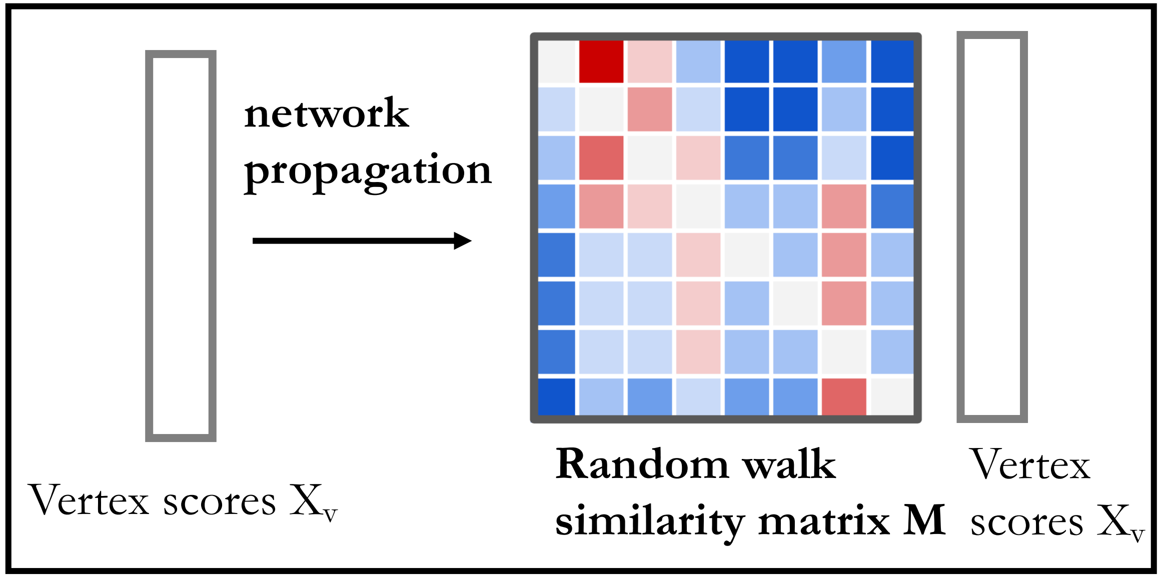
Propagation family

$\mathcal{S} = \mathcal{M}_{\delta,p}$: Subgraphs S with $\underline{M_{u,v} \geq \delta}$ for p fraction of $(u,v) \in S$

Vertices are "close"
via random walk

(also require $M_{v,u} \geq \delta$ if M is not symmetric, eg personalized PageRank)

In RECOMB 2022 paper: some theory and simulations show propagation family approximates subnetworks found by network propagation methods



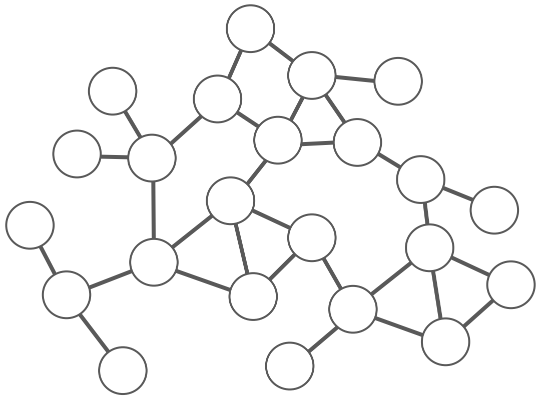
Vertex scores X_v

network propagation

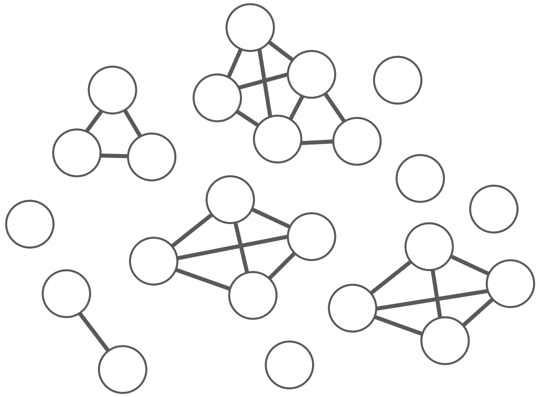
Random walk similarity matrix M

Vertex scores X_v

Alternatively: edge-dense subnetworks of *"similarity threshold graph"*

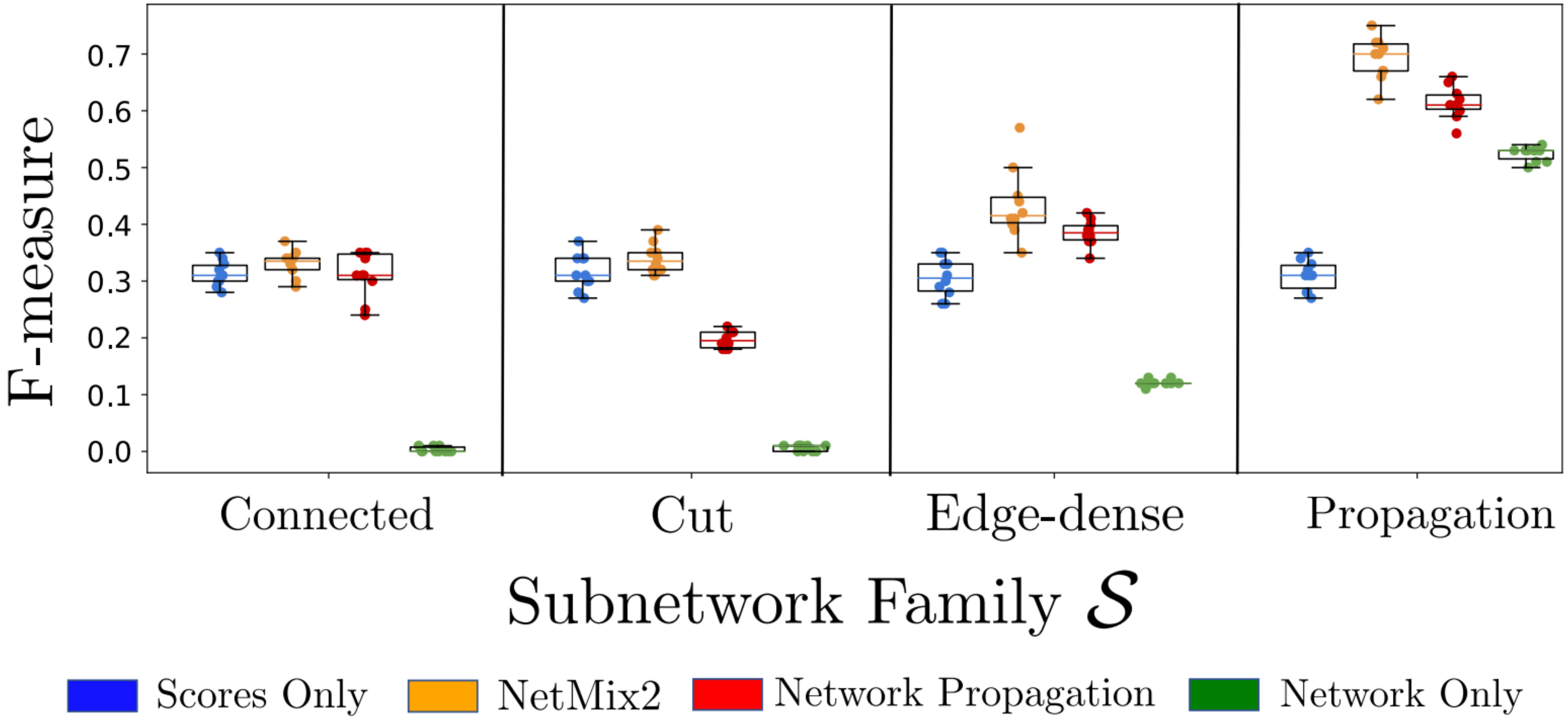


Interaction network G



Similarity threshold graph G_δ

Simulations: Propagation family corresponds to the subnetworks identified by network propagation



Scores only = {vertices w/ top-|A| scores}

Network only = {vertices w/ top-|A| vertex centrality}

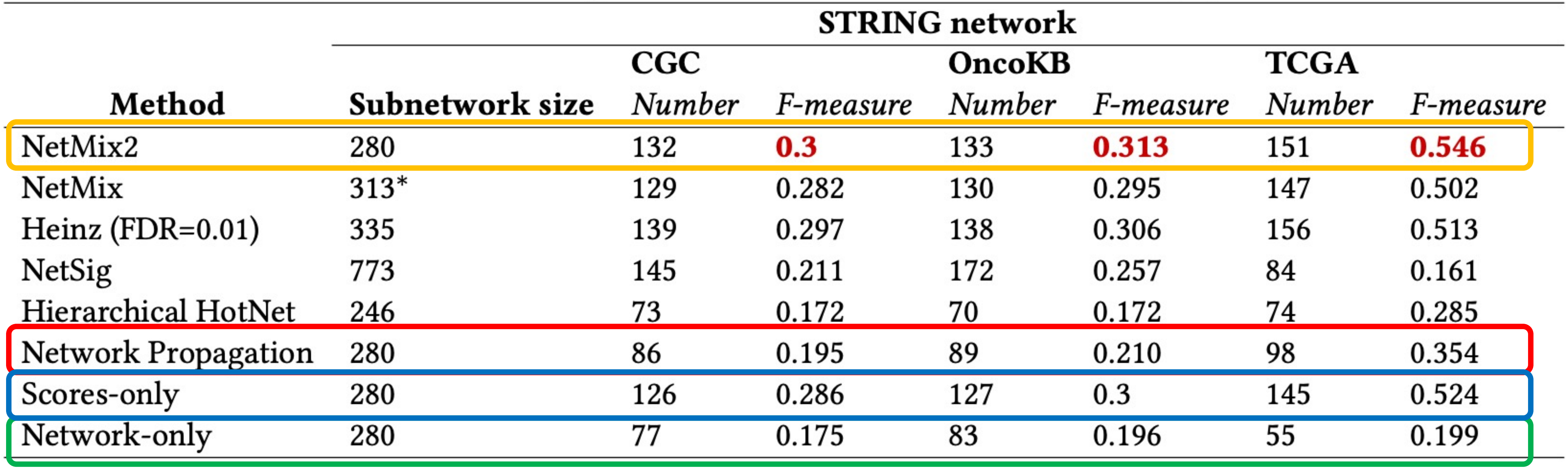
Network propagation = {vertices w/ top-|A| propagated scores}

G = HINT+HI interaction network with |G|≈15000 nodes (Leiserson et al 2015)

Altered subnetwork A of size |A|=0.01n selected uniformly at random from subnetwork family $\mathcal{S}$

Results: somatic mutations in cancer

NetMix2 outperforms other methods at identifying previously reported driver mutations in cancer.



		STRING network					
		CGC		**OncoKB**		**TCGA**	
Method	**Subnetwork size**	*Number*	*F-measure*	*Number*	*F-measure*	*Number*	*F-measure*
NetMix2	280	132	**0.3**	133	**0.313**	151	**0.546**
NetMix	313*	129	0.282	130	0.295	147	0.502
Heinz (FDR=0.01)	335	139	0.297	138	0.306	156	0.513
NetSig	773	145	0.211	172	0.257	84	0.161
Hierarchical HotNet	246	73	0.172	70	0.172	74	0.285
Network Propagation	280	86	0.195	89	0.210	98	0.354
Scores-only	280	126	0.286	127	0.3	145	0.524
Network-only	280	77	0.175	83	0.196	55	0.199

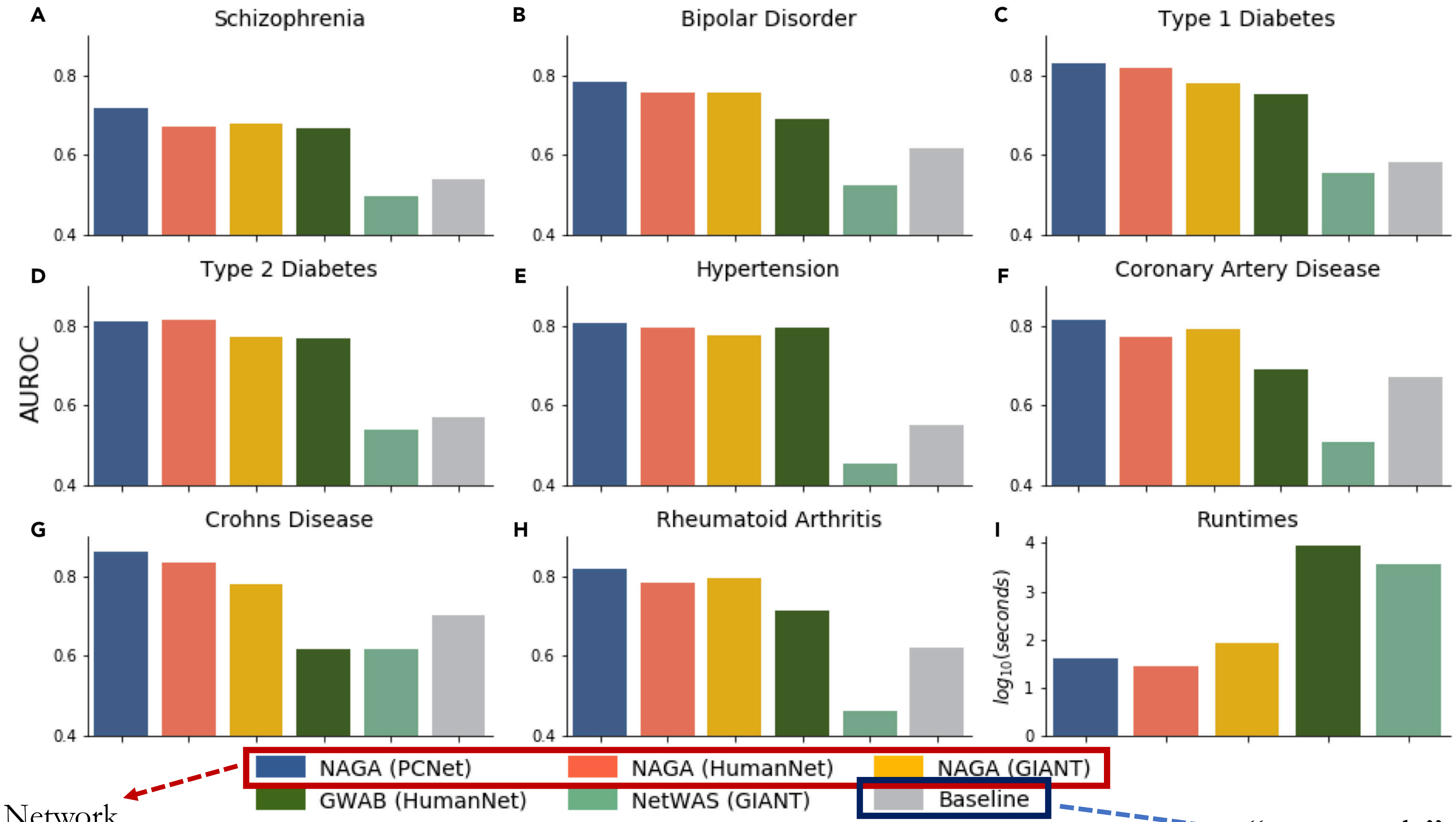
G = STRING protein interaction network

Vertex scores X_v = MutSIg2CV z-scores computed based on frequency of somatic mutations in TCGA tumor samples

Note: "Scores-only" has good performance – how helpful is interaction network?

Results: GWAS

Recent study by Carlin et al (iScience 2019) – evaluates how well methods identify known disease reference genes

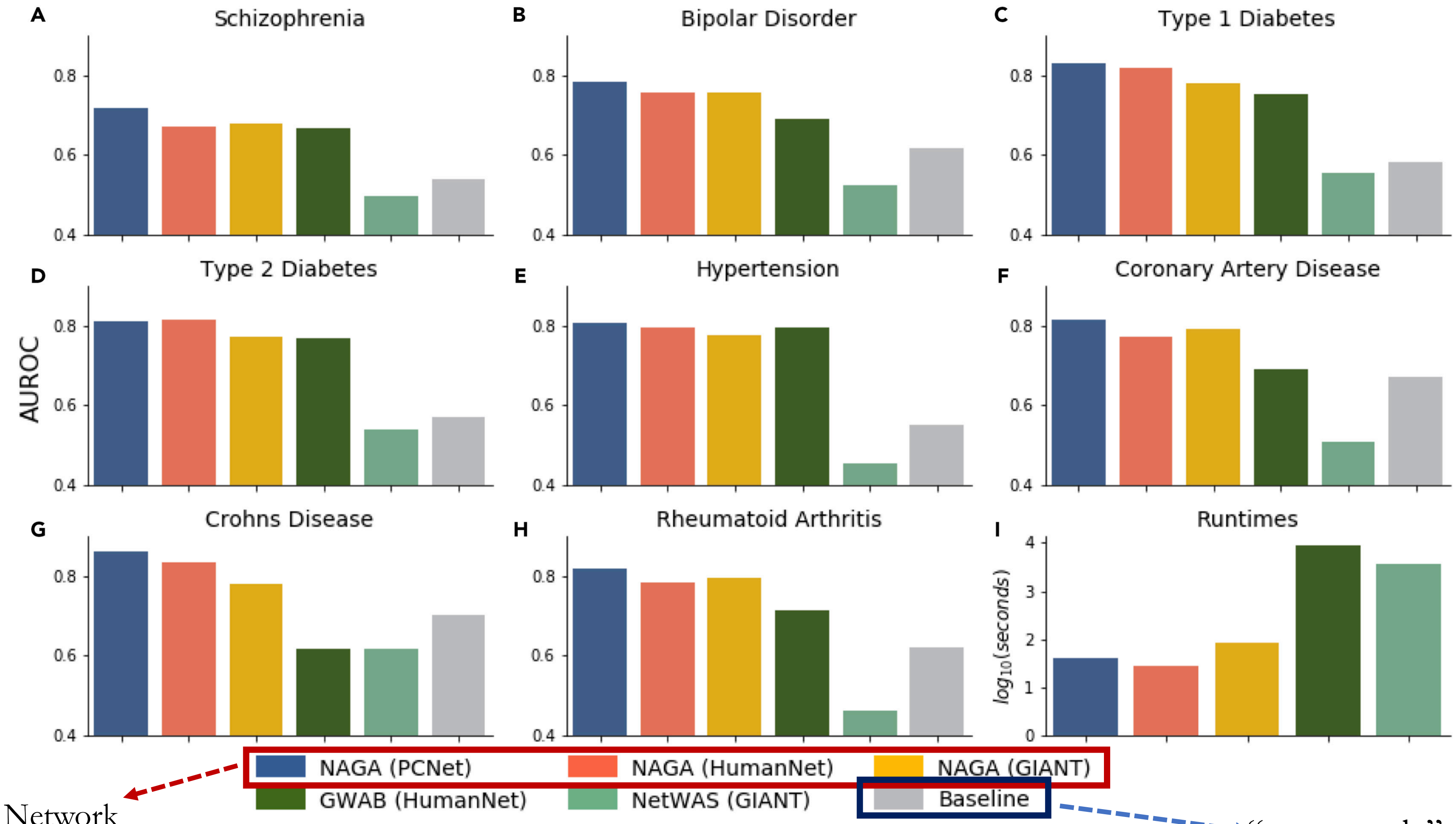


Network propagation

“scores only”

Results: GWAS

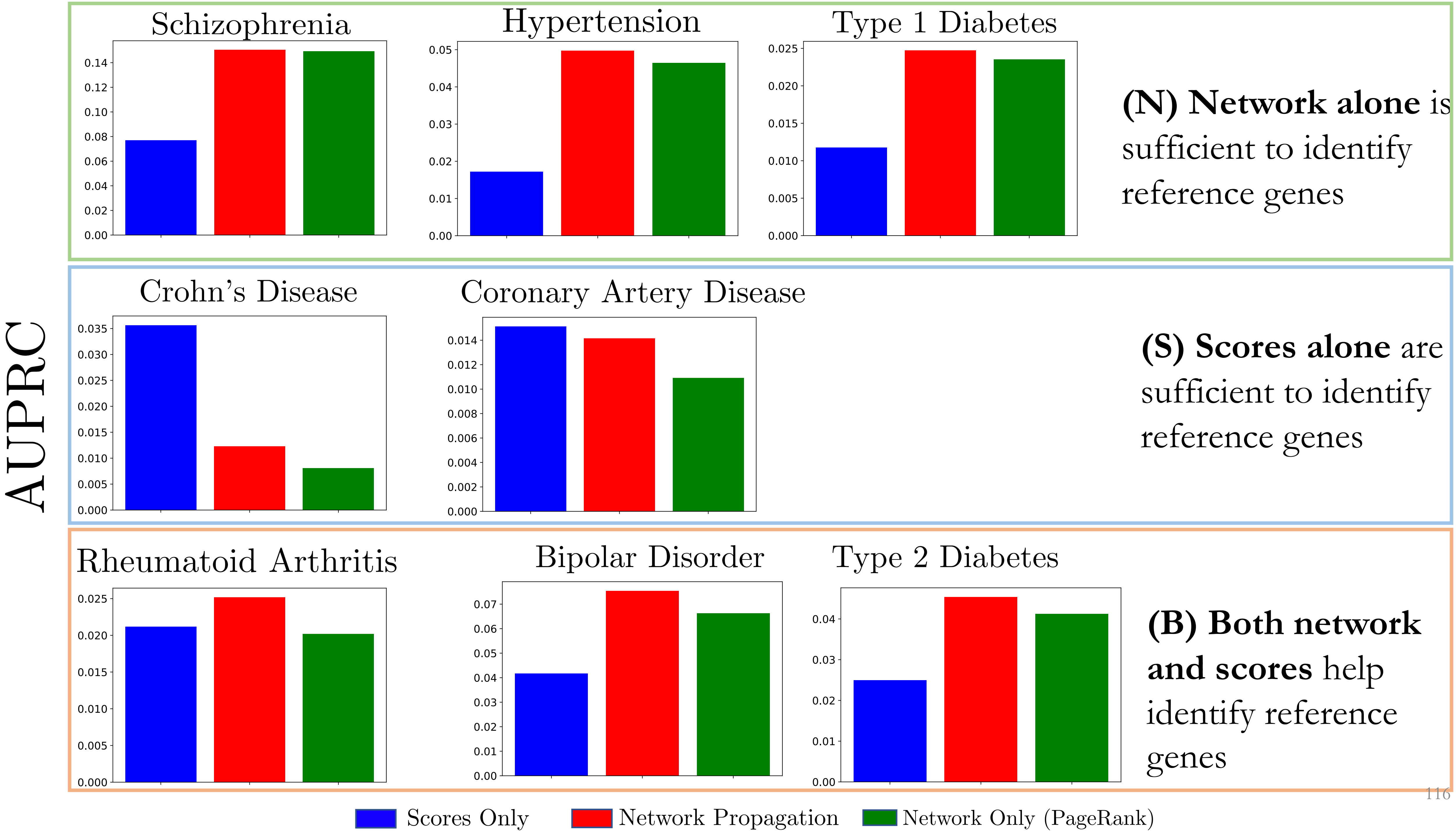
Recent study by Carlin et al (iScience 2019) – evaluates how well methods identify known disease reference genes



Network propagation

“scores only”

Issue: AUROC is poor metric for small reference sets! (<1% of 15,000 genes)

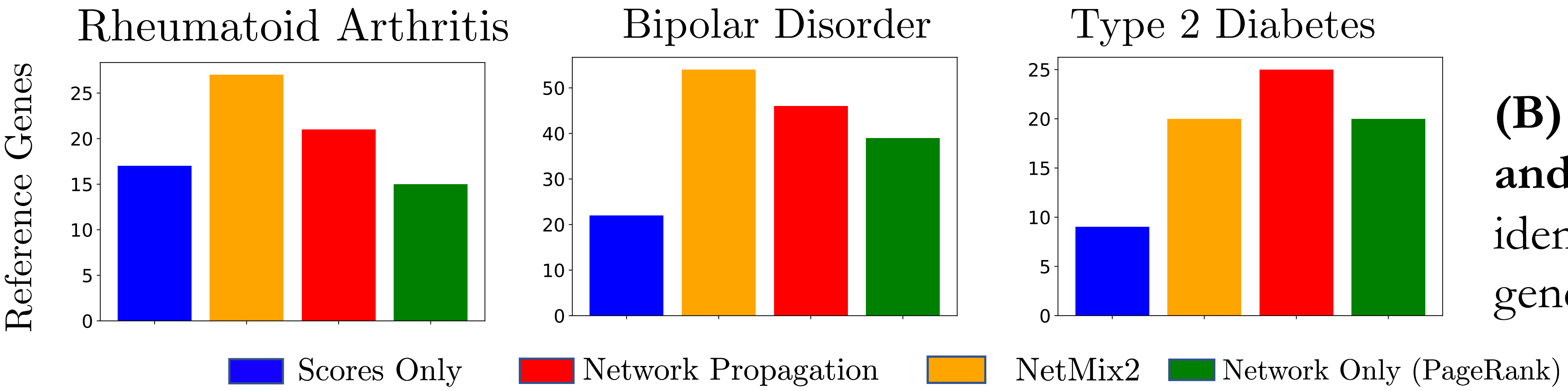


(N) Network alone is sufficient to identify reference genes

(S) Scores alone are sufficient to identify reference genes

(B) Both network and scores help identify reference genes

NetMix2 results on diseases where both network and scores help



(B) Both network and scores help identify reference genes

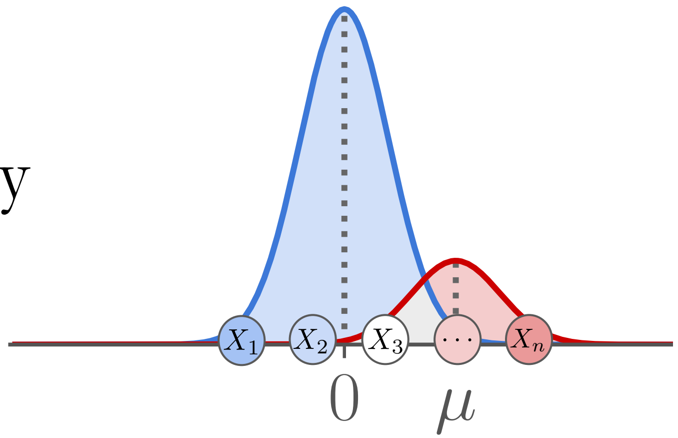
NetMix2 outperforms network propagation on 2/3 diseases

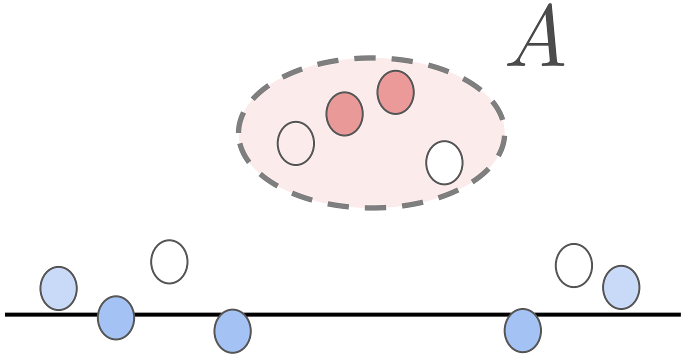
Anomaly detection

Normal means: Data $X_1, \ldots, X_n$ independently distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0,1) & \text{otherwise} \end{cases}$$

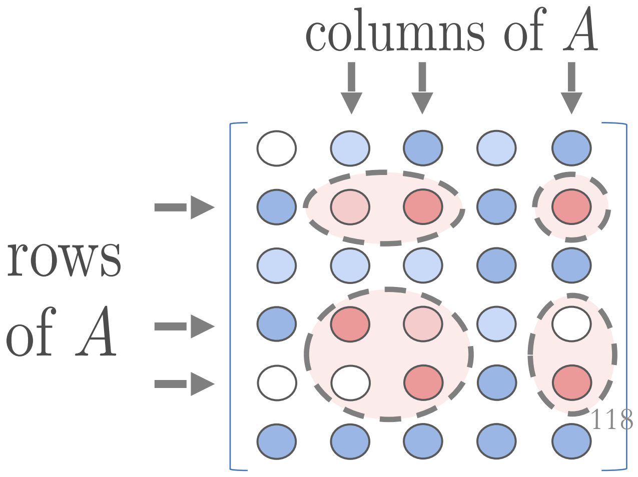
for anomaly A





Long history in statistics/ML:

- Unstructured anomalies: *Localfdr/empirical Bayes* methods (e.g. Efron et al., JASA 2001/2004, *Annals of Stats* 2007, etc), *Higher criticism* (Donoho and Jin, *Annals of Stats* 2004, etc), …
- Structured anomalies
    - **Intervals:** Jeng et al (JASA 2010)
    - **Submatrices:** Kolar et al (NeurIPS 2011), Chen and Xu (ICML 2014), Brennan et al (COLT 2018), Liu and A-C (KDD 2019)
    - **Connected subgraphs:** Qian et al (NeurIPS 2014), Aksoylar et al (ICML 2017), Cadena et al (AAAI 2018/TKDD 2019)
    - **Subgraphs w/ small cut**: Sharpnack et al (NeurIPS 2013/AISTATS 2013)
    - **Other:** Brennan et al (ICML 2020)



Generalizing to anomaly detection

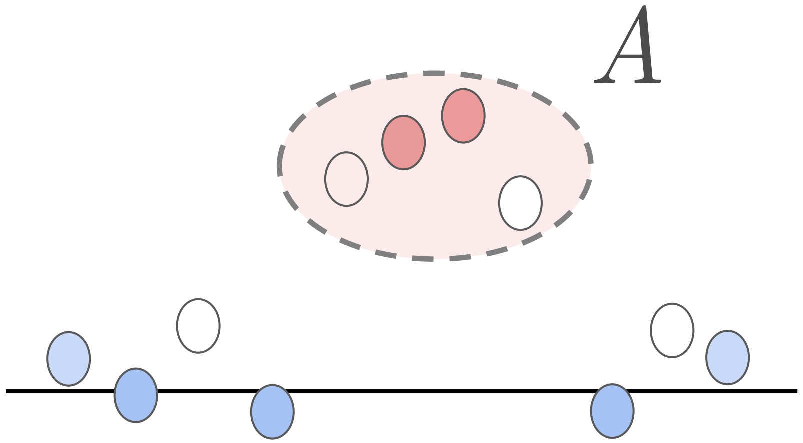
- $\mathcal{S}$ is **anomaly family** (set of subsets of {1, …, n})
- $A \in \mathcal{S}$ is the anomaly

Examples of anomaly families:

Interval family

$\mathcal{S} = \mathcal{I}_n$ = intervals {i, i+1, …, j}

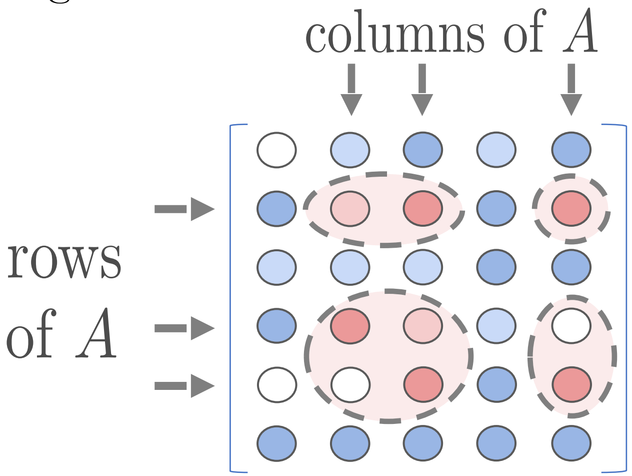
Changepoint detection



Submatrix family

$\mathcal{S} = \mathcal{M}_N$ = submatrices of N

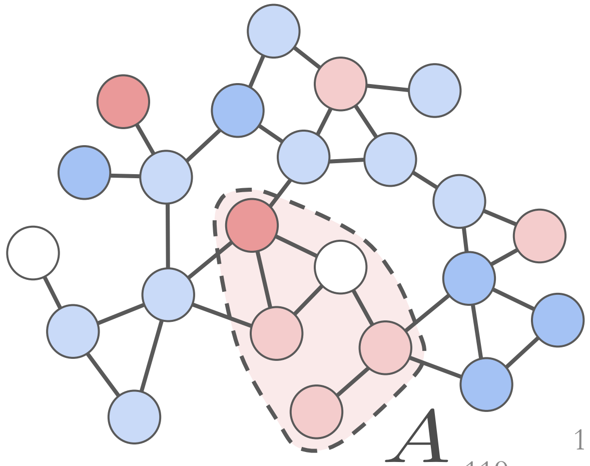
Bi-clustering



Connected family

$\mathcal{S} = \mathcal{C}_G$ = connected subgraphs of G

Network anomaly detection



Anomalous Subset ~~Altered Subnetwork~~ Distribution

- n datapoints (e.g. vertices of interaction network)
- $\mathcal{S}$ is **anomaly family** (set of subsets of {1, …, n})
- $A \in \mathcal{S}$ is the anomaly

Datapoints $(X_1, \ldots, X_n)$ distributed as $X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0,1) & \text{otherwise} \end{cases}$

Anomalous Subset Problem (ASP): Given data$(X_1, \ldots, X_n)$ and anomaly family $\mathcal{S}$, find anomaly A.

Maximum Likelihood Estimator (MLE):

$$\widehat{A}_{\text{MLE}} = \arg\max_{S \in \mathcal{S}} \frac{1}{\sqrt{|S|}} \sum_{i \in S} X_i$$

MLE is optimal for some anomaly families but not others

- Jeng et al (JASA 2010) show (asymptotic) "*near-optimality*" for interval family $\mathcal{S} = \mathcal{I}_n$

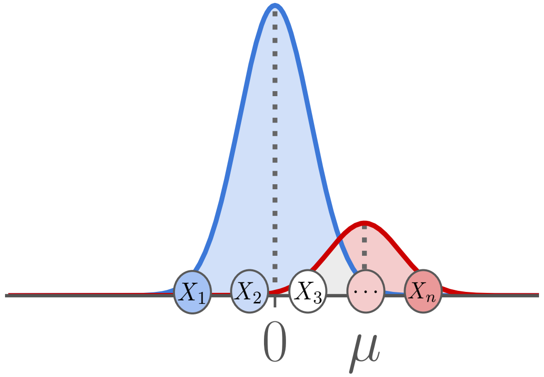
- Liu and A-C (KDD 2019) show similar guarantees for submatrix family $\mathcal{S} = \mathcal{M}_N$

But we showed that MLE is a **biased** estimator for the connected family $\mathcal{S} = \mathcal{C}_G$

Question: for which anomaly families $\mathcal{S}$ is MLE biased?

Maximum Likelihood Estimator (MLE):

$$\widehat{A}_{\text{MLE}} = \arg\max_{S \in \mathcal{S}} \frac{1}{\sqrt{|S|}} \sum_{i \in S} X_i$$



Data $X_1, \ldots, X_n$ distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$

Our contribution

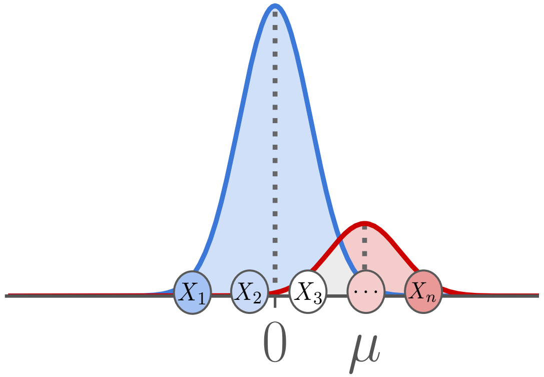
Question: For which anomaly families $\mathcal{S}$ is the MLE biased?

We show: MLE is biased <--> number of sets in anomaly family $\mathcal{S}$ that contain the anomaly A is exponential

Generalizes previous results on interval/submatrix family, which have sub-exponential size

Forward direction: ICML 2021 paper
Reverse direction: proved by Henri Schmidt+UC (unpublished)

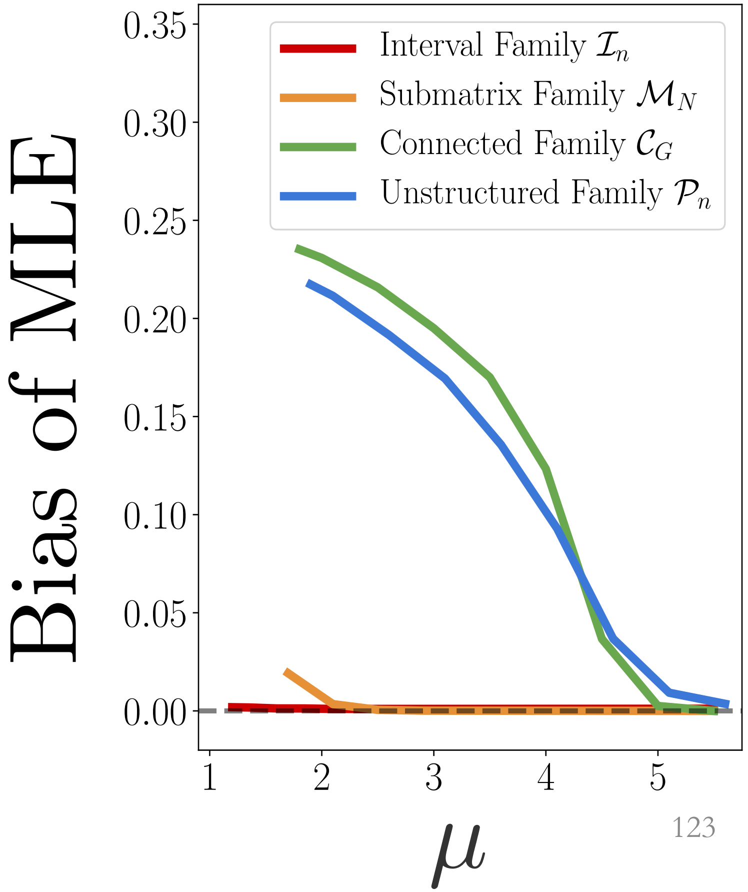
Conjecture: result holds for exponential family distr. besides normal



Data $X_1, \ldots, X_n$ distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

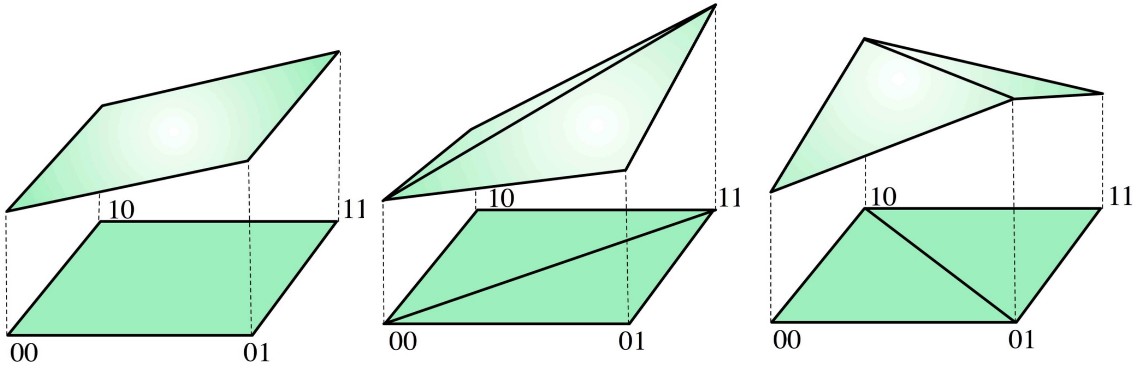
where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$

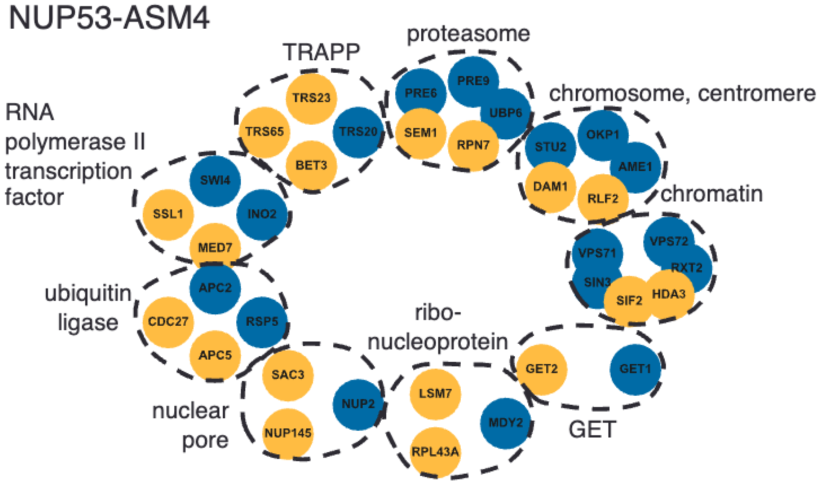


Learning genetic interactions (epistasis)

Brian Arnold

Ben Raphael





Chitra*, Arnold*, Raphael. *In review at Nature Genetics.*

* indicates joint first authorship

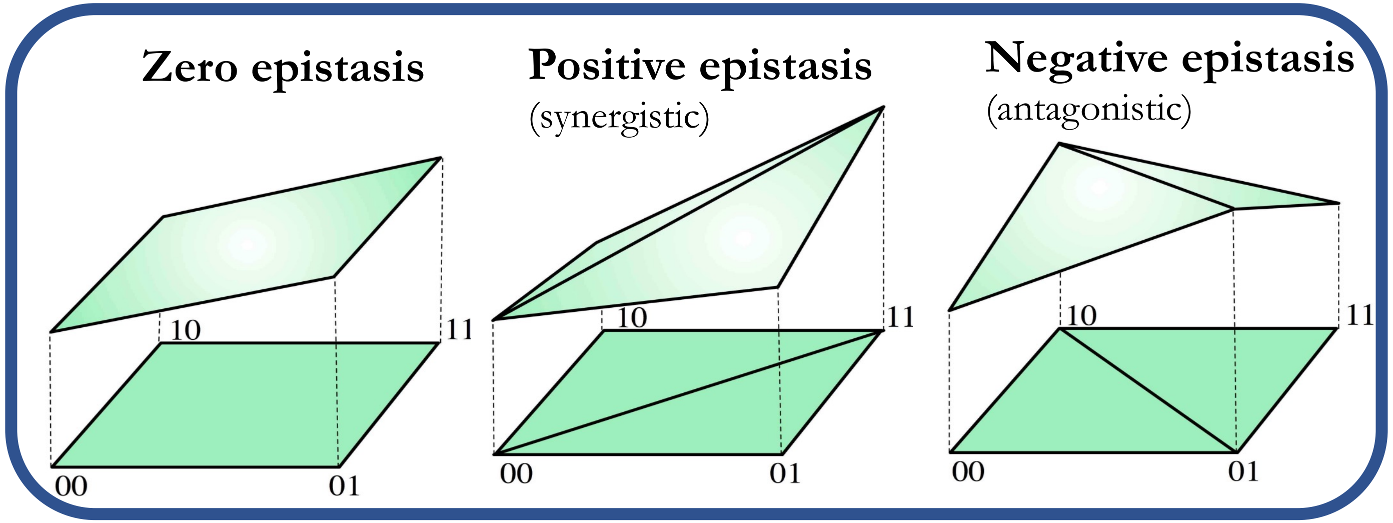
Epistasis = genetic interactions - one gene mutation changes effect of other gene mutations

Quantifying pairwise epistasis

(2 mutations)

Additive: $\epsilon = f_{11} - (f_{01} + f_{10})$

Multiplicative: $\epsilon = \dfrac{f_{11}}{f_{01} f_{10}}$



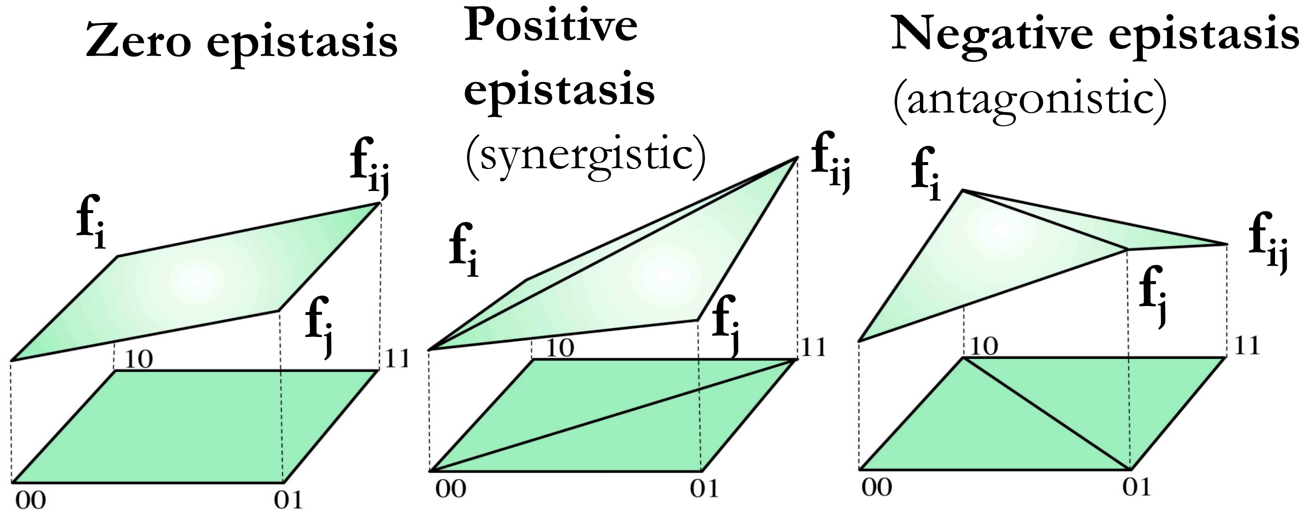
Epistasis = genetic interactions - one gene mutation changes effect of other gene mutations

Quantifying pairwise epistasis

(2 mutations)

Additive: $\epsilon = f_{ij} - f_i - f_j$

Multiplicative: $\epsilon = \dfrac{f_{ij}}{f_i f_j}$



Zero epistasis

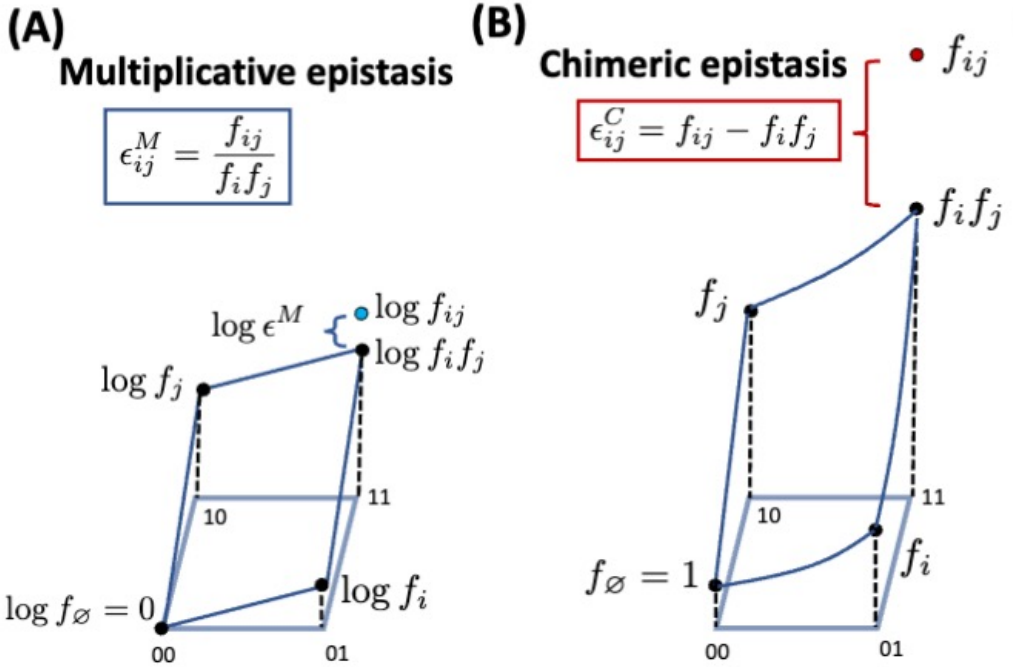
Positive epistasis (synergistic)

Negative epistasis (antagonistic)

Many papers in genetics claim to use multiplicative model but measure epistasis additively:

$$\epsilon^C = f_{ij} - f_i f_j$$

- *"Chimeric"* formula: a chimera of additive, multiplicative scales
- OK in practice: has same sign as multiplicative formula



(A) Multiplicative epistasis

$\boxed{\epsilon^M_{ij} = \dfrac{f_{ij}}{f_i f_j}}$

(B) Chimeric epistasis

$\boxed{\epsilon^C_{ij} = f_{ij} - f_i f_j}$

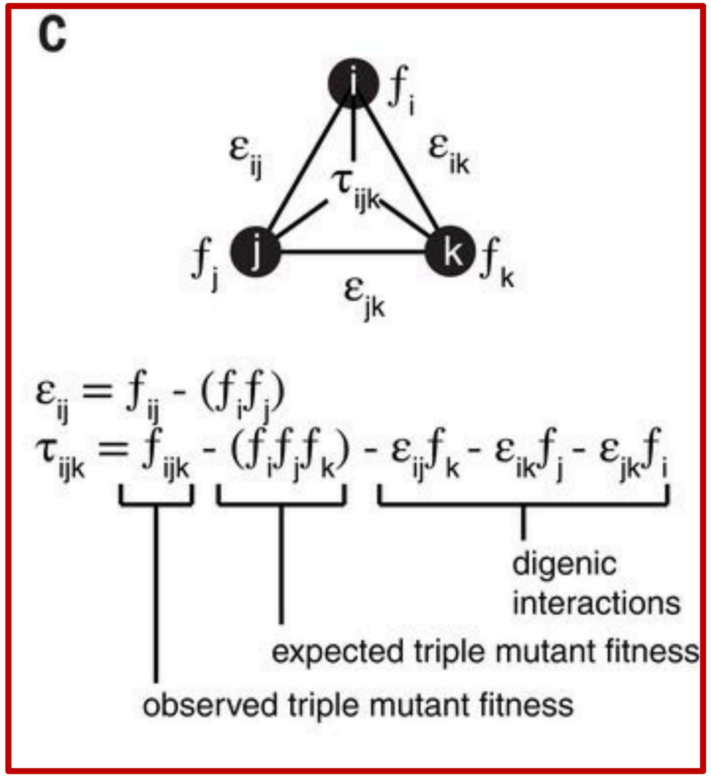
Higher-order epistasis (3+ mutations)

<u>Additive</u>:

$$\begin{aligned}\epsilon^A_{ijk} &= f_{ijk} - [f_i + f_j + f_k + \epsilon^A_{ij} + \epsilon^A_{ik} + \epsilon^A_{jk}] \\ &= f_{ijk} - f_{ij} - f_{ik} - f_{jk} + f_i + f_j + f_k.\end{aligned}$$

<u>Multiplicative</u>:

$$\epsilon^M_{ijk} = \frac{f_{ijk}}{f_i f_j f_k \epsilon^M_{ij} \epsilon^M_{ik} \epsilon^M_{jk}} = \frac{f_{ijk} f_i f_j f_k}{f_{ij} f_{jk} f_{ik}}$$

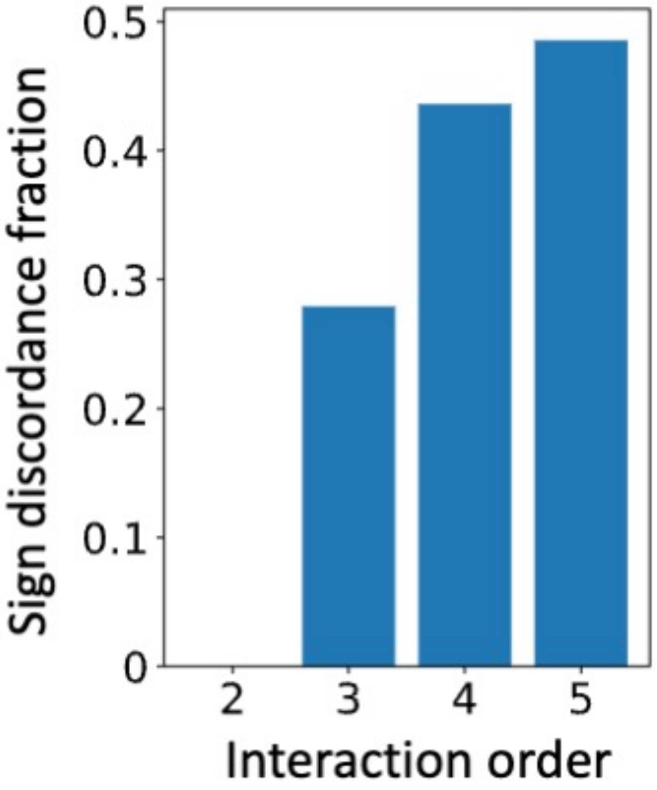


$$\varepsilon_{ij} = f_{ij} - (f_i f_j)$$

$$\tau_{ijk} = \underbrace{f_{ijk}}_{\text{observed triple mutant fitness}} - \underbrace{(f_i f_j f_k)}_{\text{expected triple mutant fitness}} \underbrace{- \varepsilon_{ij} f_k - \varepsilon_{ik} f_j - \varepsilon_{jk} f_i}_{\text{digenic interactions}}$$

Recent studies (*Science* 2018 + 2020) claim to use multiplicative fitness model but…

- Derive *"chimeric"* 3-way formula that combines additive, mult. Scales
- <u>No guarantees</u>: may have different sign versus multiplicative formula

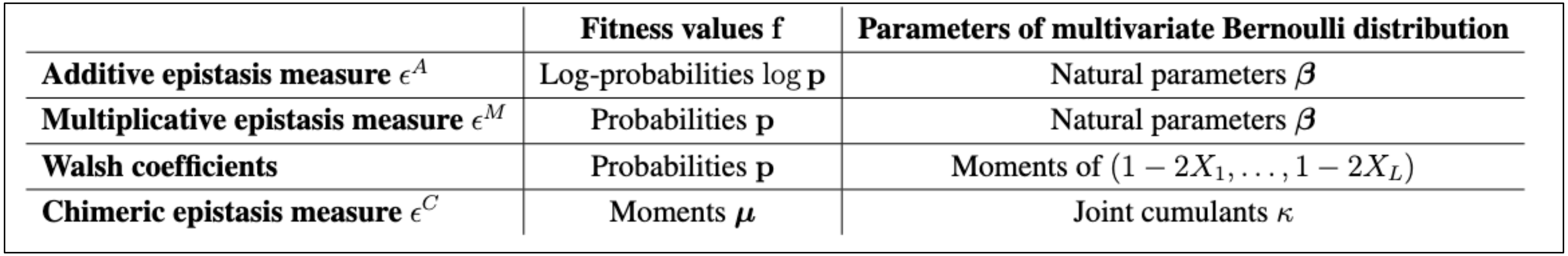


<u>Hard to trust reported interactions!</u>

Our contributions

1. Unify different epistasis formulas using probabilistic framework

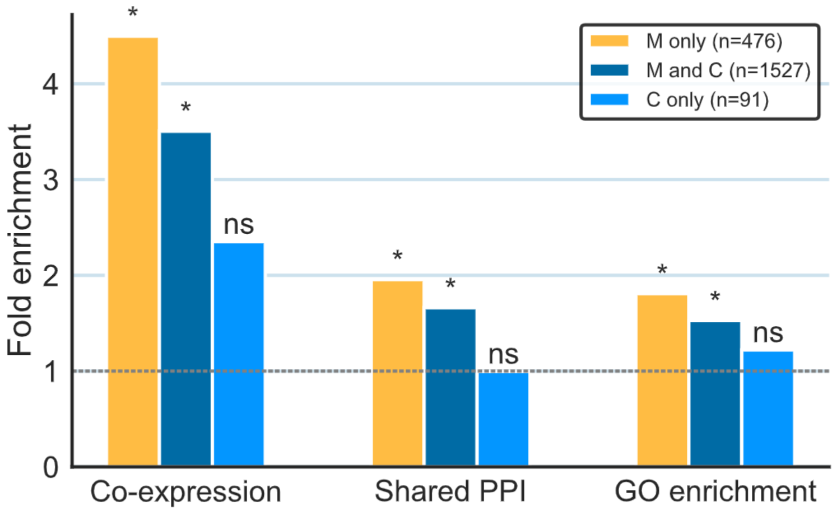
Epistasis formulas = different parametrizations of *multivariate Bernoulli* distribution (MVB)



	Fitness values f	Parameters of multivariate Bernoulli distribution
Additive epistasis measure ϵ^A	Log-probabilities $\log \mathbf{p}$	Natural parameters $\boldsymbol{\beta}$
Multiplicative epistasis measure ϵ^M	Probabilities $\mathbf{p}$	Natural parameters $\boldsymbol{\beta}$
Walsh coefficients	Probabilities $\mathbf{p}$	Moments of $(1-2X_1,\ldots,1-2X_L)$
Chimeric epistasis measure ϵ^C	Moments $\boldsymbol{\mu}$	Joint cumulants κ

Our theory shows additive/multiplicative formulas are **more statistically sound** than chimeric formulas

2. Reanalyze *Science* data – learning 3-way interactions in yeast – using correct formula



Negative *(antagonistic)* 3-way interactions = functional redundancy

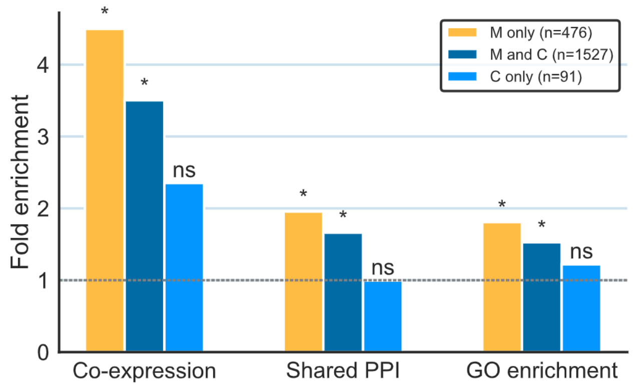
Using correct (mult.) formula finds ~500 more neg. interactions

- Significantly enriched for functional similarity measures
- **extends trigenic interaction network by 25%**

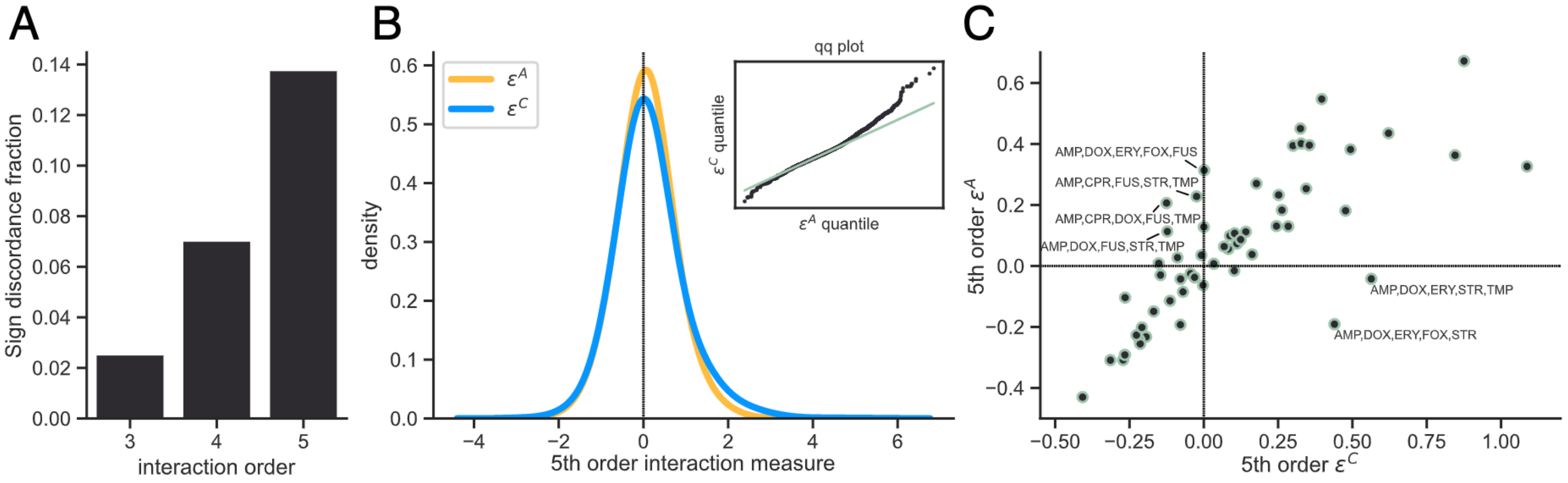
Sign disagreement leads to different biological findings

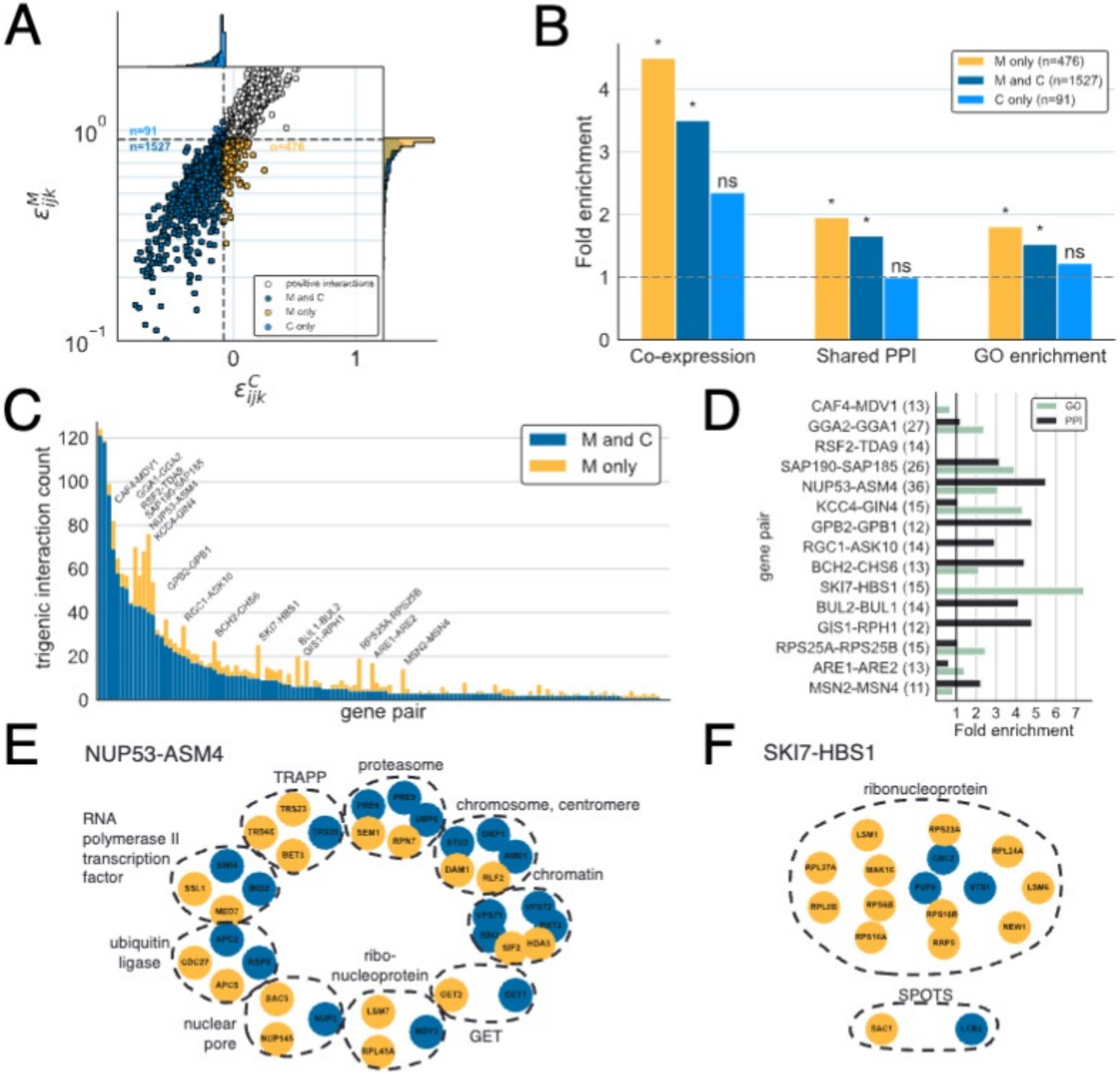
3-way epistasis in yeast

		Chimeric measure ϵ_{ijk}^{C}		
		Positive	Ambiguous	Negative
Multiplicative measure ϵ_{ijk}^{M}	Positive	1197	**259**	**0**
	Ambiguous	**116**	4291	**91**
	Negative	**10**	**466**	1527



Multi-way drug interactions





Reanalysis of trigenic yeast interactions from Kuzmin et al. (*Science* 2018/2020)